Polarization of the Rényi Information Dimension for Single and Multi Terminal Analog Compression

Saeid Haghighatshoar*, Emmanuel Abbe†
*EPFL, Lausanne, Switzerland, saeid.haghighatshoar@epfl.ch
†Princeton University, Princeton, NJ, USA, eabbe@princeton.edu

***Abstract*—This paper shows that the Rényi information dimension (RID) of an i.i.d. sequence of mixture random variables polarizes to the extremal values of 0 and 1 (fully discrete and continuous distributions) when transformed by an Hadamard matrix. This provides a natural counter-part over the reals of the entropy polarization phenomenon over finite fields. It is further shown that the polarization pattern of the RID is equivalent to the BEC polarization pattern, which admits a closed form expression. These results are used to construct universal and deterministic partial Hadamard matrices for analog to analog (A2A) compression of random i.i.d. signals. In addition, a framework for the A2A compression of multiple correlated signals is developed, providing a first counter-part of the Slepian-Wolf coding problem in the A2A setting.**

***Index Terms*—Rényi information dimension, Polarization, Information preserving matrices, Analog compression, Distributed analog compression, Compressed sensing.**

I. Introduction

A. Analog to analog compression

Analog to analog (A2A) compression of signals has recently gathered interest in information theory [12]–[15]. In A2A compression, a high dimensional analog signal $x^n \in \mathbb{R}^n$ is encoded into a lower dimensional analog signal $y^m = f_n(x^n) \in \mathbb{R}^m$. The goal is to design the encoding so as to preserve in y^m all the information about x^n, and to obtain successful decoding for a given distortion measure like MSE or error probability. In particular, the encoding may be corrupted by noise. It is worth mentioning that when the alphabet of x and y is finite, this framework falls into traditional topics of information theory such as lossless and lossy data compression, or joint source-channel coding. The novelty of A2A compression is to consider x and y to be real valued and to impose regularity constrains on the encoder, in particular linearity, as motivated by compressed sensing [1], [2].

The challenge and practicality of A2A compression is to obtain dimensionality reduction, i.e., $m/n \ll 1$, by exploiting a prior knowledge on the signal. This may be sparsity as in compressed sensing. For k-sparse signals, and without any stability or complexity considerations, it is not hard to see that the dimensionality reduction can be of order k/n. A measurement rate of order $k/n \log(n/k)$ has been shown to be sufficient to obtain stable recovery by solving tractable optimization algorithms like convex programming (l_1 minimization). This remarkable achievement has gathered tremendous amount of attention with a large variety of algorithmic solutions deployed over the past years. The vast majority of the research has however capitalized on a common sparsity model.

Several works have explored connections between information theory and compressed sensing[1], in particular [6]–[11], however it is only recently [12] that a foundation of A2A compression has been developed, shifting the attention to probabilistic signal models beyond the sparsity structure. It is shown in [12] that under linear encoding and Lipschitz-continuous decoding, the fundamental limit of A2A compression is the Rényi information dimension (RID), a measure whose operational meaning had remained marginal in information theory until [12]. In the case of a nonsingular mixture distribution, the RID is given by the mass on the continuous part, and for the specific case of sparse mixture distributions, this gives a dimensionality reduction of order k/n. It is natural to ask whether this improvement on compressed sensing is due to potentially complex or non-robust coding strategies. [13] shows that robustness to noise is not a limitation of the framework in [12]. Two other works [14], [15] have corroborated the fact that complexity may not be a limitation either. In [14] spatially-coupled matrices are used for the encoding of the signal, leveraging on the analytical ground of spatially-coupled codes and predictions of [17]. In particular, [14] shows that the RID is achieved using approximate message passing algorithm with block diagonal Gaussian measurement matrices measurement matrices. However, the size of the blocks are increasing as the measurement rate approaches the RID. In [15], using a new entropy power inequality (EPI) for integer-valued random variables that was further developed in [16], the polarization technique was used to deterministically construct partial Hadamard matrices for encoding discrete signals over the reals. This provides a way to achieve a measurement rate of $o(n)$ for signals with a zero RID along with a stable low complexity recovery algorithm. The case of mixture distributions was however left open in [15].

This paper proposes a new approach to A2A compression by means of a polarization theory over the reals. The use of polarization techniques for sparse recovery was proposed in [18] for discrete signals, relying on coding strategies over finite fields. In this paper, it is shown that using the RID, one obtains a natural counter-part over the reals of the entropy polarization phenomenon [19], [20]. Specifically, the entropy (or source) polarization phenomenon [20] shows that transforming an i.i.d.

[1] [3]–[5] investigate LDPC coding techniques for compressed sensing

sequence of discrete random variables using an Hadamard matrix polarizes the conditional entropies to the extreme values of 0 and 1 (deterministic and maximally random distributions). We show in this paper that the RID of an i.i.d. sequence of mixture random variables also polarizes to the two extreme values 0 and 1 (discrete and continuous distributions). To get to this result, properties of the RID in vector settings and related information measures are first developed. It is then shown that the RID polarization is, as opposed to the entropy polarization, obtained with an analytical pattern. In other words, there is no need to rely on algorithms to compute the set of components which tend to 0 or 1, as this is given by a known pattern equivalent to the BEC channel polarization [19]. This is then used to obtain universal A2A compression schemes based on explicit partial Hadamard matrices. The current paper focuses on the encoding strategies and on extracting the RID without specifying the decoding strategy. Numerical simulations provide evidence that efficient message passing algorithms may be used in conjunction to the obtained encoders.

Finally, the paper extends the realm of A2A compression to a multi signal settings. Techniques of distributed compressed sensing were introduced in [23] for specific classes of sparse signal models. We provide here an information theoretic framework for general multi signal A2A compression, as a counter part of the Slepian & Wolf coding problem in source compression [24]. A measurement rate region to extract the RID of correlated signals is obtained and is shown to be tight.

B. Notations and preliminaries

The set of reals, integers and positive integers will be denoted by $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{Z}_+$ respectively. $\mathbb{N} = \mathbb{Z}_+ \backslash \{0\}$ will denote the set of strictly positive integers. For $n \in \mathbb{N}$, $[n] = \{1, 2, \dots, n\}$ denotes the sequence of integers from 1 to n. For a set S, the cardinality of the set will be denoted by $|S|$, thus $|[n]| = n$.

All random variables are denoted by capital letters and their realization by lower case letter (x is a realization of the random variable X). The expected value and the variance of a random variable X are denoted by $\mathbb{E}\{X\}$ and σ_X^2. For $i, j \in \mathbb{Z}$, X_i^j is a column vector consisting of the random variables $\{X_i, X_{i+1}, \dots, X_j\}$ and for $i > j$, we set X_i^j equal to null.

For a discrete random variable X with a distribution p_X, $H(X) = H(p_X)$ denotes the discrete entropy of X. For the continuous case, $h(X) = h(p_X)$ denotes the differential entropy of X. Throughout the paper, we assume that all of discrete and continuous random variables have well-defined discrete entropy and differential entropy respectively. For random elements X, Y and Z, $I(X;Y)$ and $I(X;Y|Z)$ denote the mutual information of X and Y and the conditional mutual information of X and Y given Z. $I(X;Y|z)$ denotes the mutual information of X and Y given a specific realization $Z = z$. Hence, $I(X;Y|Z) = \mathbb{E}_Z\{I(X;Y|z)\}$. For simplicity, we also assume that all of the random variables (discrete, continuous or mixture) have finite second order moments.

All probability distributions are assumed to be nonsingular. Hence, in the general case for a random variable X, the distribution of X can be decomposed as $p_X = \delta p_c + (1-\delta) p_d$, where p_c and p_d are the continuous and the discrete part of the distribution and $0 \leq \delta \leq 1$ is the weight of the continuous part. Thus, $\delta = 0$ and $\delta = 1$ corresponds to the fully discrete and fully continuous case respectively. For such a probability distribution, the Rényi information dimension is interchangeably denoted by $d(p_X)$ or $d(X)$ and is equal to the weight of the continuous part δ.

There is another representation for a random variable X that we will repeatedly use in the paper. Assume U is a continuous random variable with probability distribution p_c and V is a discrete random variable with probability distribution p_d and U and V are independent. Let $\Theta \in \{0, 1\}$ be a binary valued random variable, independent of U and V with $\mathbb{P}(\Theta = 1) = \delta$. It is easy to see that we can represent X as $X = \Theta U + \bar{\Theta} V$, where $\bar{\Theta} = 1 - \Theta$. In this case, the random variable X will have the distribution $p_X = \delta p_c + (1 - \delta) p_d$. Also, if X_1^n is a sequence of such random variables with the corresponding binary random variables Θ_1^n, $C_\Theta = \{i \in [n] : \Theta_i = 1\}$ is a random set consisting of the position of the continuous components of the signal. Similarly, $\bar{C}_\Theta = [n] \backslash C_\Theta$ is defined to be the position of the discrete components.

For a matrix Φ of a given dimension $m \times n$ and a set $S \subset [n]$, Φ_S is a sub-matrix of dimension $m \times |S|$ consisting of those columns of Φ having index in S. Similarly, for a vector of random variables X_1^n, the vector $X_S = \{X_i : i \in S\}$ is a sub-vector of X_1^n consisting of those random variables having index in S. For two matrices A and B of dimensions $m_1 \times n$ and $m_2 \times n$, $[A; B]$ denotes the $(m_1 + m_2) \times n$ matrix obtained by vertically concatenating A and B.

For an $x \in \mathbb{R}$ and a $q \in \mathbb{N}$, $[x]_q = \frac{\lfloor qx \rfloor}{q}$ denotes the uniform quantization of x by interspacing $\frac{1}{q}$. Similarly, for a vector of random variables X_1^n, $[X_1^n]_q$ will denote the component-wise uniform quantization of X_1^n.

For $a(q)$ and $b(q)$ two functions of q, $a(q) \preceq b(q)$ or equivalently $b(q) \succeq a(q)$ will be used for

$$\lim_{q \to \infty} \frac{b(q) - a(q)}{\log_2(q)} \geq 0.$$

Similarly, $a(q) \doteq b(q)$ is equivalent to $a(q) \preceq b(q), a(q) \succeq b(q)$.

An ensemble of single terminal measurement matrices will be denoted by $\{\Phi_N\}$, where N is the labeling sequence and can be any subsequence of $\mathbb{N}$. The dimension of the family will be denoted by $m_N \times N$, where m_N is the number of measurements taken by Φ_N. The asymptotic measurement rate of the ensemble is defined by $\limsup_{N \to \infty} \frac{m_N}{N}$. We will also work with an ensemble of multi terminal measurement matrices. We will focus to the two terminal case and the extension to more than two terminals will be straightforward. We will denote these two terminals by x, y and the corresponding ensemble by $\{\Phi_N^x, \Phi_N^y\}$ with the corresponding dimension $m_N^x \times N$ and $m_N^y \times N$. The measurement rate vector for this ensemble will be denoted by (ρ_x, ρ_y), where $\rho_x = \limsup_{N \to \infty} \frac{m_N^x}{N}, \rho_y = \limsup_{N \to \infty} \frac{m_N^y}{N}$.

II. Rényi information dimension

Let X be a random variable with a probability distribution p_X over $\mathbb{R}$. The upper and the lower RID of this random variable are defined as follows:

$$\bar{d}(X) = \limsup_{q\to\infty} \frac{H([X]_q)}{\log_2(q)},$$
$$\underline{d}(X) = \liminf_{q\to\infty} \frac{H([X]_q)}{\log_2(q)}.$$

By Lebesque decomposition or Jordan decomposition theorem, any probability distribution over $\mathbb{R}$ like p_X can be written as a convex combination of a discrete part, a continuous part and a singular part, namely,

$$p_X = \alpha_d p_d + \alpha_c p_c + \alpha_s p_s,$$

where p_d, p_c and p_s denote the discrete, continuous and the singular part of the distribution and $\alpha_d, \alpha_c, \alpha_s \geq 0$ and $\alpha_d + \alpha_c + \alpha_s = 1$. In [27], Rényi showed that if $\alpha_s = 0$, namely, there is no singular part in the distribution and $p_X = (1-\delta)\, p_d + \delta\, p_c$ for some $\delta \in [0,1]$, then the RID is well-defined and $d(X) = \bar{d}(X) = \underline{d}(X) = \delta$. Moreover, he proved that if X_1^n is a continuous random vector then $\lim_{q\to\infty} \frac{H([X_1^n]_q)}{\log_2(q)} = n$, implying the RID of n for the n-dimensional continuous random vector.

Our objective is to extend the definition of RID for arbitrary vector random variables, which are not necessarily continuous. To do so, we first restrict ourselves to a rich space of random variables with well-defined RID. Over this space, it will be possible to give a full characterization of the RID as we will see in a moment.

Definition 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a standard probability space. The space $\mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$ is defined as $\mathcal{L} = \cup_{n=1}^{\infty} \mathcal{L}_n$, where $\mathcal{L}_1$ is the set of all nonsingular random variables and for $n \in \mathbb{N}\backslash\{1\}$, $\mathcal{L}_n$ is the space of n-dimensional random vectors defined as

$$\mathcal{L}_n = \{X_1^n : \text{ there exist } k \in \mathbb{N}, A \in \mathbb{R}^{n\times k} \text{ and } Z_1^k \text{ independent and nonsingular such that } X_1^n = AZ_1^k\}.$$

Remark 1. It is not difficult to see that all n-dimensional vector random variables, singular or nonsingular, can be well approximated in the space $\mathcal{L}$, for example in ℓ_2-sense. However, this is not sufficient to fully characterize the RID. Specially, the RID is discontinuous in ℓ_p topology, $p \geq 1$. For example, we can construct a sequence of fully discrete random variables in $\mathcal{L}$ converging to a fully continuous random variable in ℓ_p, whereas the RID of the sequence is 0 and does not converge to 1. Although we have such a mathematical difficulty in giving a characterization of the RID, we think that the space $\mathcal{L}$ is rich enough for modeling most of the cases that we encounter in applications.

Over $\mathcal{L}$, we will generalize the definition of the RID to include joint RID, conditional RID and Rényi information defined as follows.

Definition 2. Let X_1^n be a random vector in $\mathcal{L}$. The joint RID of X_1^n provided that it exists, is defined as

$$d(X_1^n) = \lim_{q\to\infty} \frac{H([X_1^n]_q)}{\log_2(q)}.$$

Definition 3. Let (X_1^n, Y_1^m) be a random vector in $\mathcal{L}$. The conditional RID of X_1^n given Y_1^m and Rényi information of Y_1^n about X_1^n, provided they exist, are defined as follows:

$$d(X_1^n|Y_1^m) = \lim_{q\to\infty} \frac{H([X_1^n]_q|Y_1^m)}{\log_2(q)}$$
$$I_R(X_1^n; Y_1^n) = d(X_1^n) - d(X_1^n|Y_1^m).$$

Generally, it is difficult to give a characterization of RID for a general multi-dimensional distribution because it can contain probability mass over complicated subsets or submanifolds of lower dimension. However, we will show that the vector Rényi information dimension is well-defined for the space $\mathcal{L}$. In order to give the characterization of RID over $\mathcal{L}$, we also need to define some concepts from linear algebra of matrices, namely, for two matrices of appropriate dimensions, we propose the following definition of the "influence" of one matrix on another matrix and "residual" of one matrix given another matrix.

Definition 4. Let A and B be two arbitrary matrices of dimension $m_1 \times n$ and $m_2 \times n$. Also let $K \subset [n]$. The influence of the matrix B on the matrix A and the residual of the matrix A given B over the column set K are defined to be

$$I(A;B)[K] = \operatorname{rank}([A;B]_K) - \operatorname{rank}(A_K),$$
$$R(A;B)[K] = \operatorname{rank}([A;B]_K) - \operatorname{rank}(B_K).$$

Remark 2. It is easy to check that $I(A;B)[K]$ is the amount of increase of the rank of the matrix A_K by adding rows of the matrix B_K and $R(A;B)[K]$ is the residual rank of the matrix A_K knowing the rows of the matrix B_K. Moreover, one can easily check that $I(A;B)[K] = R(B;A)[K]$.

Theorem 1. *Let (X_1^n, Y_1^m) be a random vectors in the space $\mathcal{L}$, namely, there are i.i.d. nonsingular random variables Z_1^k and two matrices A and B of dimension $n\times k$ and $m\times k$ such that $X_1^n = AZ_1^k$ and $Y_1^m = BZ_1^k$. Let $Z_i = \Theta_i U_i + \bar{\Theta}_i V_i$ be the representation for $Z_i,\ i \in [k]$. Then, we have*

1) $d(X_1^n) = \mathbb{E}\{\operatorname{rank}(A_{C_\Theta})\}$,
2) $d(X_1^n|Y_1^m) = \mathbb{E}\{R(A;B)[C_\Theta]\}$,

where $C_\Theta = \{i \in [k] : \Theta_i = 1\}$ is the random set consisting of the position of continuous components.

Remark 3. Notice that the results intuitively make sense, namely, for a specific realization θ_1^k if $\theta_i = 0$ we can neglect Z_i because it is fully discrete and does not affect the RID. Moreover, over the continuous components the resulting contribution to the RID is equal to the rank of the matrix A_{C_θ}, which is the effective dimension of the space over which the continuous random variable $A_{C_\theta}U_{C_\theta}$ is distributed. Finally, all of these contributions are averaged over all possible realizations of Θ_1^k.

Using Theorem 1, it is possible to prove a list of properties of the RID.

Theorem 2. *Let (X_1^n, Y_1^m) be a random vector in $\mathcal{L}$ as in Theorem 1. Then, we have the following properties:*

1) $d(X_1^n) = d(MX_1^n)$ *for any arbitrary invertible matrix M of dimension $n \times n$.*
2) $d(X_1^n, Y_1^m) = d(X_1^n) + d(Y_1^m|X_1^n)$.
3) $I_R(X_1^n; Y_1^m) = I_R(Y_1^m; X_1^n)$.
4) $I_R(X_1^n; Y_1^m) \geq 0$ *and* $I_R(X_1^n; Y_1^m) = 0$ *if and only if X_1^n and Y_1^m are independent after removing discrete common parts, namely, those $Z_i, i \in [k]$ that are fully discrete.*

Further investigation also shows that we have a very nice duality between the discrete entropy and the RID as depicted in Table I. As we will see in Subsection III-B and III-C, this duality can be generalized to include some of the theorems in classical information theory like single terminal and multi terminal (Slepian & Wolf) source coding problems.

Discrete random variables	Random variables in $\mathcal{L}$
Discrete entropy H	RID d
Conditional entropy	Conditional RID
Mutual information	Rényi mutual information
Deterministic	Discrete
Chain rule	Chain rule
Single terminal source coding	Single terminal A2A compression
Multi terminal source coding	Multi terminal A2A compression

TABLE I: Duality between H and d

III. Main results

In this section, we will give a brief overview of the results proved in the paper. Subsection III-A is devoted to the results obtained for the polarization of the Rényi information dimension. These results are used in Subsections III-B and III-C to study *A2A compression* problem from an information theoretic point of view. Subsection III-B considers the single terminal case whereas Subsection III-C is devoted to the multi terminal case.

A. Polarization of the Rényi information dimension

Before stating the polarization result for the RID, we define the m-dimensional erasure process as follows.

Definition 5. Let $\alpha \in [0, 1]$. An "erasure process" with initial value α is defined as follows.

1) $e^{\emptyset} = \alpha$. $e^+ = 2\alpha - \alpha^2$ and $e^- = \alpha^2$.
2) Let $e_n = e^{b_1 b_2 \dots b_n}$, for some arbitrary $\{+, -\}$-valued sequence b_1^n. Define

$$e_n^+ = e^{b_1 b_2 \dots b_n +} = 2e_n - e_n^2,$$
$$e_n^- = e^{b_1 b_2 \dots b_n -} = e_n^2.$$

Remark 4. Notice that using the $\{+, -\}$ labeling, we can construct a binary tree where each leaf of the tree is assigned a specific $\{+, -\}$-valued sequence.

Let $\{B_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. uniform $\{+, -\}$-valued random variables. By replacing B_1^n for $\{+, -\}$-labeling b_1^n in the definition of the erasure process, we obtain a stochastic process $e_n = e^{B_1 B_2 \dots B_n}$. Let $\mathcal{F}_n$ be the σ-field generated by B_1^n. Using the BEC polarization [19], [21], we have the following results:

1) $(e_n, \mathcal{F}_n)$ is a positive bounded martingale.
2) e_n converges to $e_\infty \in \{0, 1\}$ with $\mathbb{P}(e_\infty = 1) = \alpha$.
3) For any $0 < \beta < \frac{1}{2}$, $\liminf_{n\to\infty} \mathbb{P}(e_n \leq 2^{-N^\beta}) = 1 - \alpha$, where $N = 2^n$ is the number of all possible cases that e_n can take.

Let $n \in \mathbb{N}$ and $N = 2^n$. Assume that X_1^N is a sequence of i.i.d. nonsingular random variables with a RID equal to $d(X)$ and let $Z_1^N = H_N X_1^N$, where H_N is the Hadamard matrix of order N. For $i \in [N]$, let us define $I_n(i) = d(Z_i|Z_1^{i-1})$. Assume that b_1^n is the binary expansion of $i - 1$. By replacing 0 by $+$ and 1 by $-$, we can equivalently represent $I_n(i)$ be a sequence of $\{+, -\}$ values, namely, $I_n(i) = I^{b_1 b_2 \dots b_n}$. Similar to the erasure process, we can convert I_n to a stochastic process $I_n = I^{B_1 B_2 \dots B_n}$ by using i.i.d. uniform $\{+, -\}$-valued random variables B_1^n. We have the following theorem.

Theorem 3 (Single terminal RID polarization)**.** $(I_n, \mathcal{F}_n)$ *is an erasure stochastic process with initial value $d(X)$ polarizing to $\{0, 1\}$.*

For $n \in \mathbb{N}$ and $N = 2^n$, let $\{(X_i, Y_i)\}$ be a sequences of random vectors in the space $\mathcal{L}$, with joint and conditional RID $d(X, Y)$, $d(X|Y)$ and $d(Y|X)$. Let $Z_1^N = H_N X_1^N$ and assume that $W_1^N = H_N Y_1^N$. Let us define two processes I_n and J_n as follows.

$$I_n(i) = d(Z_i|Z_1^{i-1}), i \in [N],$$
$$J_n(i) = d(W_i|W_1^{i-1}, Z_1^N), i \in [N].$$

Similarly, we can label I_n and J_n by a sequence of b_1^n and convert them to stochastic processes $I_n = I^{B_1 B_2 \dots B_n}$ and $J_n = J^{B_1 B_2 \dots B_n}$. By this definition, we have the following theorem.

Theorem 4 (Multi terminal RID polarization)**.** $(I_n, \mathcal{F}_n)$ *and* $(J_n, \mathcal{F}_n)$ *are erasure stochastic processes with initial value $d(X)$ and $d(Y|X)$, both polarizing to $\{0, 1\}$.*

Remark 5. In the t terminal case $t > 2$ for a t terminal source $(X_1, X_2, \dots X_t)$, using a similar method it is possible to construct erasure processes with initial values $d(X_1), d(X_2|X_1), \dots, d(X_t|X_1^{t-1})$, polarizing to $\{0, 1\}$.

B. Single terminal A2A compression

In this subsection, we will use the properties of the RID developed in Section II to study the A2A compression of memoryless sources. We assume that we have a memoryless source with some given probability distribution. The idea is to capture the information of the source, to be made clearer in a moment, by taking some linear measurements. As is usual in information theory, we are mostly interested in asymptotic regime for large block lengths. To do so, we will use an

ensemble of measurement matrices to analyze the asymptotic behavior. We will also define the notion of REP (restricted isoentropy property) for an ensemble of measurement matrices. This subsection is devoted to the single terminal case. The results for the multi terminal case will be given in Subsection III-C. We are mostly interested to the the measurement rate region of the problem in order to successfully capture the source.

Definition 6. Let X_1^N be a sequence of i.i.d. random variables with a probability distribution p_X (discrete, mixture or continuous) over $\mathbb{R}$, and let $D_1^N = [X_1^N]_q$ for $q \in \mathbb{N}$. The family of measurement matrices $\{\Phi_N\}$, indexed with a subsequence of $\mathbb{N}$ and with dimension $m_N \times N$, is ϵ-REP(p_X) with the measurement rate ρ if

$$\limsup_{q\to\infty} \frac{H(D_1^N|\Phi_N X_1^N)}{H(D_1^N)} \leq \epsilon, \qquad (1)$$
$$\limsup_{N\to\infty} \frac{m_N}{N} = \rho.$$

To give some intuitive justification for the REP definition, let us assume that all of the measurements are captured with a device with finite precision $\frac{1}{q_0}$ for some $q_0 \in \mathbb{N}$. In that case, although the potential information of the signal, in terms of bits, can be very large, but what we effectively observe through the finite precision device is only $H([X_1^N]_{q_0})$. In such a setting, the ratio of the information we lose after taking the measurements, assuming that some genie gives us the infinite precision measurement captured from the signal, is exactly what we have in the definition of REP, namely,

$$\frac{H(D_1^N|\Phi_N X_1^N)}{H(D_1^N)}, \qquad (2)$$

where we assume that $D_1^N = [X_1^N]_{q_0}$. This might be a reasonable model for application because pretty much this is what happens in reality. The problem with this model is that it is not invariant under some obvious transformations like scaling. For example, assume that we are scaling the signal by some real number. In this case, through some simple examples it is possible to show that the ratio in (2) can change considerably. There are two approaches to cope with this problem. One is to scale the signal with a desired factor to match it to the finite precision quantizer, which in its own can be very interesting to analyze but probably will be two complicated. The other way, is to take our approach and develop a theory for the case in which the resolution is high enough so that the quality measure proposed in (2) is not affected by the shape of the distribution of the signal.

Remark 6. Notice that in the fully discrete case, the REP definition is simplified to the equivalent form

$$\frac{H(X_1^N|\Phi_N X_1^N)}{H(X_1^N)} \leq \epsilon,$$
$$\limsup_{N\to\infty} \frac{m_N}{N} \leq \rho.$$

Remark 7. For a non discrete source with strictly positive RID, $d(X) > 0$, if we divide the numerator and the denumerator in the expression (1) by $\log_2(q)$, take the limit as q tends to infinity and use the definition of the RID, we get the equivalent form

$$\frac{d(X_1^N|\Phi_N X_1^N)}{d(X_1^N)} \leq \epsilon.$$

Interestingly, this implies that in the high resolution regime that we are considering for analysis, the information isometry (keeping more than $1-\epsilon$ ratio of the information of the signal) is equivalent to the Rényi isometry. Moreover, from the properties of the RID, it is easy to see that this REP measure meets some of the invariance requirements that we expect. For example, it is scale invariant and any invertible linear transformation of the input signal X_1^N keeps the ϵ-REP measure unchanged.

We can also extend the definition when the probability distribution of the source is not known exactly but it is known to belong to a given collection of distributions Π.

Definition 7. Assume $\Pi = \{\pi : \pi \in \Pi\}$ is a class of nonsingular probability distributions over $\mathbb{R}$. The family of measurement matrices $\{\Phi_N\}$, indexed with a subsequence of $\mathbb{N}$ and with dimension $m_N \times N$, is ϵ-REP(Π) for measurement rate ρ if it is ϵ-REP(π) for every $\pi \in \Pi$.

Now that we have the required tools and definitions, we give a characterization of the required measurement rate in order to keep the information isometry. Similar to all theorems in information theory, we do this using the "converse" and "achievability" parts.

Theorem 5 (Converse result)**.** *Let X_1^N be a sequence of i.i.d. random variables in $\mathcal{L}$. Suppose $\{\Phi_N\}$ is a family of ϵ-REP(p_X) measurement matrices of dimension $m_N \times N$, then $\rho \geq d(X_1)(1-\epsilon)$.*

Remark 8. This result implies that to capture the information of the signal the asymptotic measurement rate must be approximately greater then the RID of the source. This in some sense is similar to the single terminal source coding problem in which the encoding rate must be grater then the entropy of the source. This again the emphasizes the analogy between H and d. Moreover, in the discrete case, $d(X) = 0$, the result is trivial.

Remark 9. It was proved in [12] that under linear encoding and block error probability distortion condition, the measurement rate must be higher than the RID of the source, $\rho \geq d(X)$. Theorem 5 strengthen this result stating that $\rho \geq d(X)$ must hold even under the milder ϵ-REP restriction on the measurement ensemble.

Theorem 5 puts a lower bound on the measurement rate in order to keep the ϵ-REP property. However, it might happen that there is no measurement family to achieve this bound. Fortunately, as we will see, it is possible to deterministically truncate the family of Hadamard matrices to obtain a measurement family with ϵ-REP property and measurement rate $d(X)$. This is summarized in the following two theorems. Notice that in the fully continuous case as Theorem 5 implies, the feasible

measurement rate is approximately 1 which for example can be achieved with any complete orthonormal family, thus no explicit construction is necessary. For the noncontinuous case, we will distinguish between the fully discrete case and the mixture case because they need different proof techniques. Theorem 6 and 7 summarize the results.

Theorem 6 (Achievability result)**.** *Let X_1^N be a sequence of i.i.d. discrete integer[2]-valued random variables. Then, for any $\epsilon > 0$, there is a family of ϵ-REP(p_X) partial Hadamard matrices of dimension $m_N \times N$, for $N = 2^n$ with $\rho = 0$.*

Theorem 7 (Achievability result)**.** *Let X_1^N be a sequence of i.i.d. random variables in $\mathcal{L}$. Then, for any $\epsilon > 0$, there is a family of ϵ-REP(p_X) partial Hadamard matrices of dimension $m_N \times N$, for $N = 2^n$ with $\rho = d(X_1)$.*

We have also the general result in Theorem 8 which implies that we can construct a family of truncated Hadamard matrices which is ϵ-REP for a class of distributions.

Theorem 8 (Achievability result)**.** *Let Π be a family of probability distributions with strictly positive RID. Then, for any $\epsilon > 0$, there is a family of ϵ-REP(Π) partial Hadamard matrices of dimension $m_N \times N$, for $N = 2^n$, with $\rho = \sup_{\pi \in \Pi} d(\pi)$.*

Remark 10. Theorem 8 implies that there is a fixed ensemble of measurement matrices capable of capturing the information of the all of the distributions in the family Π. This is very useful in applications because usually taking the measurements is costly and most of the time we do not have the exact distribution of the signal. If each distribution needs its own specific measurement matrix, we have to do several rounds of the measurement each time taking the measurements compatible with one specific distribution and do the recovery process for that specific distribution. The benefit of Theorem 8 is that one measurement ensemble works for all of distributions. It is also good to notice that although the measurement ensemble is fixed, the recovery (decoding) process might need to know the exact distribution of the signal in order to have successful recovery.

C. Multi terminal A2A compression

In this section, our goal is to extend the A2A compression theory from the single terminal case to the multi terminal case. In the multi terminal setting, we have a memoryless source which is distributed in more than one terminal and we are going to take linear measurements from different terminals in order to capture the information of the source. We are again interested in an asymptotic regime for large block lengths. To do so, we will use an ensemble of distributed measurement matrices that we will introduce in a moment. Similar to the single terminal case, we are interested in the measurement rate region of the problem, namely, the number of measurements that we need from different terminals in order to capture the signal faithfully. We will analyze the problem for two terminal case. The extension to more than two terminals is straightforward.

Definition 8. Let $\{(X_i, Y_i)\}_{i=1}^N$ be a two terminal memoryless source with (X_1, Y_1) being in $\mathcal{L}$. The family of distributed measurement matrices $\{\Phi_N^x, \Phi_N^y\}$, indexed with a subsequence of $\mathbb{N}$, is ϵ-REP$(p_{X,Y})$ for the measurement rate (ρ_x, ρ_y) if

$$\limsup_{q\to\infty} \frac{H([X_1^N]_q, [Y_1^N]_q | \Phi_N^x X_1^N, \Phi_N^y Y_1^N)}{H([X_1^N]_q, [Y_1^N]_q)} \le \epsilon, \quad (3)$$
$$\limsup_{N\to\infty} \frac{m_N^x}{N} \le \rho_x, \quad \limsup_{N\to\infty} \frac{m_N^y}{N} \le \rho_y.$$

Remark 11. If (X, Y) is a random vector in $\mathcal{L}$ with $d(X, Y) > 0$, similar to what did in the single terminal case, dividing the numerator and the denumerator in the expression (3) by $\log_2(q)$ and taking the limit as q tends to infinity, we get the equivalent definition

$$\frac{d(X_1^N, Y_1^N | \Phi_N^x X_1^N, \Phi_N^y Y_1^N)}{d(X_1^N, Y_1^N)} \le \epsilon,$$

which implies the equivalence of the information isometry and the Rényi isometry.

Remark 12. Notice that in the fully discrete case, the definition above is simplified to the equivalent form

$$\frac{H(X_1^N, Y_1^N | \Phi_N^x X_1^N, \Phi_N^y Y_1^N)}{H(X_1^N, Y_1^N)} \le \epsilon,$$
$$\limsup_{N\to\infty} \frac{m_N^x}{N} \le \rho_x, \quad \limsup_{N\to\infty} \frac{m_N^y}{N} \le \rho_y.$$

We can also extend the definition to a class of probability distributions.

Definition 9. Assume that $\Pi = \{\pi : \pi \in \Pi\}$ is a class of nonsingular probability distributions in $\mathcal{L}$. The family of measurement matrices $\{\Phi_N^x, \Phi_N^y\}$ is ϵ-REP(Π) for measurement rate (ρ_x, ρ_y) if it is ϵ-REP(π) for every $\pi \in \Pi$.

Definition 10. Let (X, Y) be a two dimensional random vector in $\mathcal{L}$ with a distribution $p_{X,Y}$. The Rényi information region of $p_{X,Y}$ is the set of all $(\rho_x, \rho_y) \in [0, 1]^2$ satisfying

$$\rho_x \ge d(X|Y), \ \rho_y \ge d(Y|X), \ \rho_x + \rho_y \ge d(X, Y).$$

Definition 11. Assume that Π is a class of two dimensional random vectors from $\mathcal{L}$. The Rényi information region of the class Π is the intersection of the Rényi information regions of the distributions in Π.

Similar to the single terminal case, we are interested in the rate region of the problem. We have the following converse and achievability results.

Theorem 9 (Converse result)**.** *Let $\{(X_i, Y_i)\}_{i=1}^N$ be a two-terminal memoryless source with (X_1, Y_1) being in $\mathcal{L}$. Assume*

[2]We proved this theorem using the EPI result we developed in [16], where we proved the result for lattice discrete random variables. However, we believe that such a result is also true for non-lattice discrete distributions.

that the distributed family of measurement matrices $\{\Phi_N^x, \Phi_N^y\}$ is ϵ-REP with a measurement rate (ρ_x, ρ_y). Then,

$$\rho_x + \rho_y \geq d(X,Y)(1-\epsilon),$$
$$\rho_x \geq d(X|Y) - \epsilon d(X,Y),\ \rho_y \geq d(Y|X) - \epsilon d(X,Y).$$

Remark 13. This rate region is very similar to the rate region of the distributed source coding (Slepian & Wolf) problem with the only difference that the discrete entropy has been replaced by the RID, which again emphasizes the analogy between the discrete entropy and the RID. Similar to the Slepian & Wolf problem, we call $\rho_x + \rho_y = d(X,Y)$ the dominant face of the measurement rate region.

Theorem 10 (Acievability result)**.** *Let $\{(X_i, Y_i)\}_{i=1}^N$ be a discrete two-terminal memoryless source. Then there is a family of ϵ-REP partial Hadamard matrices $\{\Phi_N^x, \Phi_N^y\}$ with $(\rho_x, \rho_y) = (0,0)$.*

Theorem 11 (Achievability result)**.** *Let $\{(X_i, Y_i)\}_{i=1}^N$ be a two-terminal memoryless source with (X_1, Y_1) belonging to $\mathcal{L}$. Given any (ρ_x, ρ_y) satisfying*

$$\rho_x + \rho_y \geq d(X_1, Y_1), \rho_x \geq d(X_1|Y_1), \rho_y \geq d(Y_1|X_1),$$

there is a family of ϵ-REP partial Hadamard matrices with measurement rate (ρ_x, ρ_y).

We have also the general result in Theorem 12 which implies that we can construct a family of truncated Hadamard matrices which is ϵ-REP for a class of distributions.

Theorem 12 (Achievability result)**.** *Let Π be a family of two dimensional probability distributions in $\mathcal{L}$. Then, for any (ρ_x, ρ_y) in the measurement region of Π, there is a family of partial Hadamard matrices which is ϵ-REP (Π) with a measurement rate (ρ_x, ρ_y).*

IV. Proof techniques

In this section, we will give a brief overview of the techniques used to prove the results. We will divide this section into three subsections. In Subsection IV-A, we will overview the proof techniques for the RID. Subsection IV-C and IV-D will be devoted to proof ideas and intuitions about the A2A compression problem in the single and multi terminal case.

A. Rényi information dimension

in this section we will prove Theorem 1 and 2 and we will give further intuitions about the RID over the space $\mathcal{L}$.

Proof of Theorem 1: To prove the first part of the theorem, notice that

$$H([X_1^n]_q) \doteq H([X_1^n]_q, \Theta_1^k) \doteq H([X_1^n]_q|\Theta_1^k),$$

because $H(\Theta_1^k) \leq k \doteq 0$. As $\Theta_1^k \in \{0,1\}^k$ and takes finitely many values, it is sufficient to show that for any realization θ_1^k,

$$\lim_{q\to\infty} \frac{H([X_1^n]_q|\theta_1^k)}{\log_2(q)} = \operatorname{rank}(A_{C_\theta}). \tag{4}$$

Taking the expectation over Θ_1^k, we will get the result. To prove (4), notice that

$$\begin{aligned} H([X_1^n]_q|\theta_1^k) &= H([A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q) \\ &\doteq H([A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q|V_{\bar{C}_\theta}) \quad (5) \\ &\doteq H([A_{C_\theta}U_{C_\theta}]_q), \quad (6) \end{aligned}$$

where we used $H(V_{\bar{C}_\theta}) \leq NH(V_1) \doteq 0$. We also used the fact that knowing $V_{\bar{C}_\theta}$, $[A_{C_\theta}U_{C_\theta}]_q$ and $[A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q$ are equal up to finite uncertainty. Specifically, suppose L is the minimum number of lattices of size $\frac{1}{q}$ required to cover $A_{\bar{C}_\theta} \times [0, \frac{2}{q}]^{|\bar{C}_\theta|}$, which is a finite number. Then

$$H([A_{C_\theta}U_{C_\theta}]_q|V_{\bar{C}_\theta}, [A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q) \leq \log_2(L),$$

which implies (5) and (6).

Generally A_{C_θ} is not full rank. Assume that the rank of A_{C_θ} is equal to m and let A_m be a subset of linearly independent rows. It is not difficult to see that knowing $[A_m U_{C_\theta}]_q$ there is only finite uncertainty in the remaining components of $[A_{C_\theta}U_{C_\theta}]_q$, which is negligible compared with $\log_2(q)$ as q tends to infinity. Therefore, we obtain

$$\begin{aligned} H([X_1^n]_q|\theta_1^k) &\doteq H([A_{C_\theta}U_{C_\theta}]_q) \\ &\doteq H([A_m U_{C_\theta}]_q) \\ &\doteq m\log_2(q). \end{aligned}$$

Thus, taking the limit as q tends to infinity, we obtain

$$\lim_{q\to\infty} \frac{H([X_1^n]|\theta_1^k)}{\log_2(q)} = \operatorname{rank}(A_{C_\theta}).$$

Also, taking the expectation with respect to Θ_1^k, we obtain $d(X_1^n) = \mathbb{E}\{\operatorname{rank}(A_{C_\Theta})\}$, which is the desired result.

To prove the second part of the theorem, notice that

$$H([X_1^n]_q|Y_1^m) \doteq H([X_1^n]_q|Y_1^m, \Theta_1^k).$$

For a specific realization θ_1^k we have

$$\begin{aligned} &H([X_1^n]_q|Y_1^m, \theta_1^k) \\ &\quad= H([A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q|B_{C_\theta}U_{C_\theta} + B_{\bar{C}_\theta}V_{\bar{C}_\theta}) \\ &\quad\doteq H([A_{C_\theta}U_{C_\theta} + A_{\bar{C}_\theta}V_{\bar{C}_\theta}]_q|B_{C_\theta}U_{C_\theta} + B_{\bar{C}_\theta}V_{\bar{C}_\theta}, V_{\bar{C}_\theta}) \\ &\quad\doteq H([A_{C_\theta}U_{C_\theta}]_q|B_{C_\theta}U_{C_\theta}). \end{aligned}$$

Generally, A_{C_θ} is not full-rank. Let A_m be the set of all linearly independent rows of A_{C_θ} of size m. Then

$$H([A_{C_\theta}U_{C_\theta}]_q|B_{C_\theta}U_{C_\theta}) \doteq H([A_m U_{C_\theta}]_q|B_{C_\theta}U_{C_\theta}).$$

It may happen that some of the rows of A_m can be written as a linear combination of rows of B_{C_θ}. Let A_r be the remaining matrix after dropping $m-r$ predictable rows of A_m. Given, $B_{C_\theta}U_{C_\theta}$, $A_r U_{C_\theta}$ has a continuous distribution thus

$$H([A_r U_{C_\theta}]_q|B_{C_\theta}U_{C_\theta}) \doteq r\log_2(q).$$

It is easy to check that r is exactly $R(A;B)[C_\theta]$. Therefore, taking the expectation with respect to Θ_1^k, we get

$$d(X_1^n|Y_1^m) = \mathbb{E}\{R(A;B)[C_\Theta]\}.$$

We also get the following corollary, which shows the additive property of the RID for the independent random variables from $\mathcal{L}$.

Corollary 1. *Let X_1^n be independent random variables from $\mathcal{L}$. Then $d(X_1^N) = \sum_{i=1}^N d(X_i)$.*

Proof: Notice that we can simply write $X_1^N = I_N \times X_1^N$, where I_N is the identity matrix of order N. Therefore, by the rank characterization for the RID, we have

$$d(X_1^N) = \mathbb{E}\{\mathrm{rank}(I_N[C_\Theta])\} = \mathbb{E}\{\sum_{i=1}^N \Theta_i\} = \sum_{i=1}^N d(X_i),$$

where we used the fact that the columns of I_N are linearly independent thus adding a column increases the rank by 1. Therefore, the rank of $I_N(C_\Theta)$ is equal to the number of 1's is Θ_1^N, namely, $\sum_{i=1}^N \Theta_i$. ∎

Using the results of Theorem 1, we can prove Theorem 2.

Proof of Theorem 2: For part 1, the proof is simple by considering the rank characterization. We know that $X_1^n = AZ_1^k$ and $d(X_1^n) = \mathbb{E}\{\mathrm{rank}(A_{C_\Theta})\}$. Moreover, $MX_1^n = MAZ_1^k$ thus $d(X_1^n) = \mathbb{E}\{\mathrm{rank}(MA_{C_\Theta})\}$. As M is invertible $\mathrm{rank}(A_{C_\Theta}) = \mathrm{rank}(MA_{C_\Theta})$, thus we get the result.

For part 2, notice that for any realization θ_1^k and the corresponding set C_θ,

$$\begin{aligned}\mathrm{rank}([A;B]_{C_\theta}) &= \mathrm{rank}(A_{C_\theta}) + R(B;A)[C_\theta] \\ &= \mathrm{rank}(B_{C_\theta}) + R(A;B)[C_\theta].\end{aligned}$$

Taking the expectation over Θ_1^k, we get the desired result

$$d(X_1^n, Y_1^m) = d(X_1^n) + d(Y_1^m|X_1^n) = d(Y_1^m) + d(X_1^n|Y_1^m).$$

For part 3, using the chain rule result from part 2 and applying the definition of $I_R(X_1^n; Y_1^m)$, we get

$$I_R(X_1^n; Y_1^m) = d(X_1^n) + d(Y_1^m) - d(X_1^n, Y_1^m),$$

which shows the symmetry of I_R with respect to X_1^n and Y_1^m.

For part 4, notice that for a specific realization θ_1^k, a simple rank check shows that $R(A;B)[C_\theta] \leq \mathrm{rank}(A_{C_\theta})$. Taking the expectation over Θ_1^k, we get $d(X_1^n|Y_1^m) \leq d(X_1^n)$.

If X_1^n and Y_1^m are independent, the equality follows from the definition. For the converse part, notice that if X_1^m is fully discrete then $d(X_1^n|Y_1^m) \leq d(X_1^n) = 0$. Similarly, if Y_1^m is fully discrete then $d(Y_1^m|X_1^n) \leq d(Y_1^m) = 0$ and using the identity $d(X_1^n) - d(X_1^n|Y_1^m) = d(Y_1^m) - d(Y_1^m|X_1^n)$, we get the equality. This case is fine because after removing the discrete $Z_i, i \in [k]$, either X_1^n or Y_1^m is equal to 0, namely, a deterministic value, and the independence holds.

Assume that none of X_1^n or Y_1^m is fully discrete. Without loss of generality, let Z_1^r be the non-discrete random variables among Z_1^k and let $\tilde{X}_1^n$ and $\tilde{Y}_1^m$ be the resulting random vectors after dropping the discrete constituents, namely, we have $\tilde{X}_1^n = A_r Z_1^r$ and $\tilde{Y}_1^m = B_r Z_1^r$, where A_r and B_r are the matrices consisting of the first r columns of A and B respectively. It is easy to check that $d(X_1^n) = d(\tilde{X}_1^n)$ and $d(\tilde{X}_1^n|\tilde{Y}_1^m) = d(X_1^n|Y_1^m)$. Thus it remains to show that $\tilde{X}_1^n$ and $\tilde{Y}_1^m$ are independent. As we have dropped all of the discrete components, the resulting $\Theta_i,\ i \in [r]$ are 1 with strictly positive probability. This implies that for any realization of θ_1^n and the corresponding C_θ, $R(A_r; B_r)[C_\theta] = \mathrm{rank}(A_{r,C_\theta})$. In particular, this holds for any C_θ of size 1, namely, for any column of A_r and B_r, which implies that if A_r has a non-zero column the corresponding column in B_r must be zero and if B_r has a non-zero column then the corresponding column in A_r must be zero. This implies that $\tilde{X}_1^n$ and $\tilde{Y}_1^m$ depend on disjoint subsets of the random variables Z_1^r. Therefore, they must be independent.

B. Polarization of the RID

In this section, we will prove the polarization of the RID in the single and multi terminal case as stated in Theorem 3 and Theorem 4. The main idea is to use the recursive structure of the Hadamard matrices and the rank characterization of the RID in the space $\mathcal{L}$.

Proof of Theorem 3: For the initial value, we have $I_0(1) = d(X_1)$. Let $n \in \mathbb{N}$ and $N = 2^n$. To simplify the proof, instead of the Hadamard matrices, H, we will use shuffled Hadamard matrices, $\tilde{H}$, constructed as follows: $\tilde{H}_1 = H_1$ and $\tilde{H}_{2N}$ is constructed from $\tilde{H}_N$ as follows

$$\begin{pmatrix} \tilde{h}_1 \\ \vdots \\ \tilde{h}_N \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{h}_1 & , & \tilde{h}_1 \\ \tilde{h}_1 & , & -\tilde{h}_1 \\ \vdots & , & \vdots \\ \tilde{h}_i & , & \tilde{h}_i \\ \tilde{h}_i & , & -\tilde{h}_i \\ \vdots & , & \vdots \end{pmatrix},$$

where $\tilde{h}_i$, $i \in [N]$ denotes the i-th row of the $\tilde{H}_N$. Let X_1^n be as in Theorem 3 and let $\tilde{Z}_1^N = \tilde{H}_N X_1^N$, where H_N is replaced by $\tilde{H}_N$. Also, let $\tilde{I}_n(i) = d(\tilde{Z}_i|\tilde{Z}_1^{i-1})$, $i \in [N]$. We first prove that $\tilde{I}$ is also an erasure process with initial value $d(X_1)$ and evolves as follows

$$\begin{aligned}\tilde{I}_n(i)^+ &= \tilde{I}_{n+1}(2i-1) = 2\tilde{I}_n(i) - \tilde{I}_n(i)^2 \\ \tilde{I}_n(i)^- &= \tilde{I}_{n+1}(2i) = \tilde{I}_n(i)^2,\end{aligned}$$

where $i \in [N]$ with the corresponding $\{+,-\}$-labeling b_1^n. Also, let $\tilde{H}^{i-1}$ and $\tilde{H}^i$ denote the first $i-1$ and the first i rows of $\tilde{H}_N$. Also, let $\tilde{h}_i$ denote the i-th row of $\tilde{H}_N$. Thus, we have $\tilde{Z}_1^i = \tilde{H}^i X_1^N$ and $\tilde{Z}_1^{i-1} = \tilde{H}^{i-1} X_1^N$. As X_1^N are i.i.d. nonsingular random variables, it results that $\tilde{Z}_1^i$ belong to the space $\mathcal{L}$ generated by the X_1^N random variables. Notice that using the rank characterization for the RID over $\mathcal{L}$, we have

$$d(\tilde{Z}_i|\tilde{Z}_1^{i-1}) = \mathbb{E}\{I(\tilde{H}^{i-1}; \tilde{h}_i)[C_\Theta]\},$$

where $I(\tilde{H}^{i-1}; \tilde{h}_i)[C_\Theta] \in \{0,1\}$ is the amount of increase of rank of $\tilde{H}_{C_\Theta}^{i-1}$ by adding $\tilde{h}_i$. Now, consider the stage $n+1$, where we have the shuffled Hadamard matrix $\tilde{H}_{2N}$. Consider the row i^+ which corresponds to the row $2i-1$ of $\tilde{H}_{2N}$. Now, if we look at the first block of the new matrix, we simply notice that adding $\tilde{h}_i$ has the same effect in increasing the rank of this block as it had in $\tilde{H}_N$. A similar argument holds for the second block. Moreover, adding $\tilde{h}_i$ increases the rank of the

matrix if it increases the rank of either the first or the second block or both. Let $\mathbf{1}_i(\Theta_1^N) \in \{0,1\}$ denote the random rank increase in $\tilde{H}^{i-1}$ by adding $\tilde{h}_i$, then we have

$$\mathbf{1}_{2i-1}(\Theta_1^{2N}) = \mathbf{1}_i(\Theta_1^N) + \mathbf{1}_i(\Theta_{N+1}^{2N}) - \mathbf{1}_i(\Theta_1^N)\mathbf{1}_i(\Theta_{N+1}^{2N}).$$

Θ_1^N and Θ_{N+1}^{2N} are i.i.d. random variables and a simple check shows that $\mathbf{1}_i(\Theta_1^N)$ and $\mathbf{1}_i(\Theta_{N+1}^{2N})$ are also i.i.d.. Taking the expectation value, we obtain

$$\tilde{I}_n(i)^+ = 2\tilde{I}_n(i) - \tilde{I}_n(i)^2. \tag{7}$$

Moreover, if we denote $\tilde{W}_1^N = \tilde{H}_N X_{N+1}^{2N}$, then by the structure of $\tilde{H}_N$ it is easy to see that $\tilde{I}_n(i)^+$ and $\tilde{I}_n(i)^-$ can be written as follows:

$$\tilde{I}_n(i)^+ = d(\tilde{Z}_i + \tilde{W}_i|\tilde{Z}_1^{i-1}, \tilde{W}_1^{i-1}),$$
$$\tilde{I}_n(i)^- = d(\tilde{Z}_i - \tilde{W}_i|\tilde{Z}_i + \tilde{W}_i, \tilde{Z}_1^{i-1}, \tilde{W}_1^{i-1}).$$

Using the chain rule for the RID, we have

$$\begin{aligned}\frac{\tilde{I}_n(i)^+ + \tilde{I}_n(i)^-}{2} &= \frac{1}{2}d(\tilde{Z}_i - \tilde{W}_i, \tilde{Z}_i + \tilde{W}_i|\tilde{Z}_1^{i-1}, \tilde{W}_1^{i-1}) \\ &= \frac{1}{2}d(\tilde{Z}_i, \tilde{W}_i|\tilde{Z}_1^{i-1}, \tilde{W}_1^{i-1}) \\ &= d(\tilde{Z}_i, |\tilde{Z}_1^{i-1}) = \tilde{I}_n(i),\end{aligned}$$

which along with (7), implies that $\tilde{I}_n(i)^- = \tilde{I}_n(i)^2$. Therefore, $\tilde{I}$ evolves like an erasure process with initial value $d(X)$.

Now, notice that the only difference between H_N and $\tilde{H}_N$ is the permutation of the rows, namely, there is a row shuffling matrix B_N such that $\tilde{H}_N = B_N H_N$. It was proved in [20] that B_N and H_N commute, which implies that $\tilde{H}_N X_1^N = H_N B_N X_1^N$. However, notice that X_1^N is an i.i.d. sequence and $B_N X_1^N$ is again an i.i.d. sequence with the same distribution as X_1^N. In particular, adding or removing B_N does not change the RID values, which implies that for $Z_1^N = H_N X_1^N$ and $I_n(i) = d(Z_i|Z_1^{i-1})$, $I_n(i) = \tilde{I}_n(i)$. Therefore, I is also be an erasure process with initial value $d(X)$, which polarizes to $\{0,1\}$. ∎

Using a similar technique, we can prove Theorem 4. The main idea is that (X,Y) are correlated random variables in the space $\mathcal{L}$ and they can be written as a linear combination of i.i.d. nonsingular random variables.

Proof of Theorem 4: For the initial value, we have $I_0(1) = d(X_1)$ and $J_0(1) = d(Y_1|X_1)$. As $\{(X_i, Y_i)\}_{i=1}^N$ is a memoryless source, similar to the single terminal case, it is easy to see that I is an erasure process with initial value $d(X_1)$ and it remains to show that J is also an erasure process but with initial value $d(Y_1|X_1)$.

Let $\tilde{H}^{i-1}$, $\tilde{H}^i$ and $\tilde{h}_i$ denote the first $i-1$ rows, the first i rows, and the i-th row $\tilde{H}_N$. As $X_1, Y_1 \in \mathcal{L}$ there is a sequence of i.i.d. random variables E_1^k and two vectors a_1^k and b_1^k such that $X_1 = \sum_{i=1}^k a_i E_i$ and $Y_1 = \sum_{i=1}^k b_i E_i$. As $\{(X_i, Y_i)\}_{i=1}^N$ is memoryless, there is a concatenation of sequence of i.i.d. copies of E_1^k, $E = \{E_1^k(1), E_1^k(2), \ldots, E_1^k(N)\}$, such that

$$Z_1^N = \tilde{H}_N X_1^N = [(B_N H_N) \otimes (a_1^k)^t]E,$$
$$W_1^N = \tilde{H}_N Y_1^N = [(B_N H_N) \otimes (b_1^k)^t]E,$$

where $\otimes$ denotes the Kronecker product and $(a_1^k)^t, (b_1^k)^t$ are the transpose of the column vectors a_1^k and b_1^k. Let

$$\Gamma = \{\Theta_1, \Theta_2, \ldots, \Theta_N\} \tag{8}$$

be the random element corresponding to the Θ pattern of $E_1^k(j), j \in [N]$, where $\Theta_j \in \{0,1\}^k, j \in [N]$. Using the rank result developed for the RID, it is easy to see that for every $j \in [N]$

$$\begin{aligned}J_n(j) &= d(W_j|W_1^{j-1}, Z_1^N) \\ &= \mathbb{E}\{I([H^{j-1} \otimes (b_1^k)^t; H \otimes (a_1^k)^t]; h_j \otimes (b_1^k)^t)[C_\Gamma]\}.\end{aligned}$$

For $i \in [N]$, let $\mathbf{1}_i(\Theta_1^N) \in \{0,1\}$ denote the random increase of rank of $[H^{i-1} \otimes (a_1^k)^t]_{C_\gamma}$ by adding $h_i \otimes (a_1^k)^t$. Now, consider the stage $n+1$, where we are going to combine two copies of $\tilde{H}_N$ to construct the matrix $\tilde{H}_{2N}$. The the row i corresponding to W_i is split into two new rows i^+ and i^- which correspond to the row number $2i-1$ and the row number $2i$ of $\tilde{H}_{2N}$.

$$\begin{pmatrix} \tilde{H}_N \otimes (a_1^k)^t & , & \tilde{H}_N \otimes (a_1^k)^t \\ \tilde{H}_N \otimes (a_1^k)^t & , & -\tilde{H}_N \otimes (a_1^k)^t \\ \vdots & , & \vdots \\ \tilde{h}_{i-1} \otimes (b_1^k)^t & , & \tilde{h}_{i-1} \otimes (b_1^k)^t \\ \tilde{h}_{i-1} \otimes (b_1^k)^t & , & -\tilde{h}_{i-1} \otimes (b_1^k)^t \\ \tilde{h}_i \otimes (b_1^k)^t & , & \tilde{h}_i \otimes (b_1^k)^t \end{pmatrix}$$

Similar to the single terminal case, we see that adding $\tilde{h}_i \otimes (b_1^k)^t$ increases the rank of the matrix if it increases the rank of the either the first or the second block. In other words,

$$\mathbf{1}_{2i-1}(\Theta_1^{2N}) = \mathbf{1}_i(\Theta_1^N) + \mathbf{1}_i(\Theta_{N+1}^{2N}) - \mathbf{1}_i(\Theta_1^N)\mathbf{1}_i(\Theta_{N+1}^{2N}),$$

where $\mathbf{1}_i(\Theta_1^N), \mathbf{1}_i(\Theta_{N+1}^{2N}) \in \{0,1\}$ are the corresponding amount of increase of the rank of the first and second block by adding the i-the row. In particular, Θ_1^N and $\tilde{\Theta}_1^N$ are i.i.d. so are $\mathbf{1}_i(\Theta_1^N)$ and $\mathbf{1}_i(\Theta_{N+1}^{2N})$. Taking the expectation, similar to what did in the single terminal case, we obtain that

$$J_n(i)^+ = 2J_n(i) - J_n(i)^2. \tag{9}$$

Moreover, one can also show that for $i \in [N]$,

$$\frac{J_n(i)^+ + J_n(i)^-}{2} = J_n(i),$$

which together with (9), implies that $J_n(i)^- = J_n(i)^2$. Therefore, J is also an erasure process with initial value $d(Y|X)$. Similar to the single terminal case, one can also show that the permutation matrix B_N is not necessary, thus the proof is complete. ∎

C. Single terminal A2A compression

In this part, we will overview the techniques used to prove the achievability part. The converse part, given in Theorem 5, has been proved in Appendix A. We will give separate constructions for the fully discrete case and the mixture case although the proof techniques used are very similar.

Achievability proof for the mixture case: We will give an explicit construct of the the measurement ensemble as follows. Let $n \in \mathbb{N}$ and let $N = 2^n$. Assume that X_1^N is a sequence of i.i.d. nonsingular random variables with RID equal to $d(X)$. Let $Z_1^N = H_N X_1^N$, where H_N is the Hadamard matrix of order N. Also assume that $I_n(i) = d(Z_i|Z_1^{i-1})$, $i \in [N]$. As we proved in Theorem 3, I is an erasure process with initial value $d(X)$. We will construct the measurement matrix Φ_N by selecting all of the rows of H_N with the corresponding I_n value greater than $\epsilon\, d(X)$. Therefore, we can construct the measurement ensemble $\{\Phi_N\}$ labelled with all N that are a power of 2. Assume that the dimension of Φ_N is $m_N \times N$. It remains to prove that the ensemble $\{\Phi_N\}$ is ϵ-REP with measurement rate $d(X)$. This will complete the proof of Theorem 7.

Proof of Theorem 7: We first show that the family $\{\Phi_N\}$ has measurement rate $d(X)$. Notice that the process I_n converges almost surely. Thus, it also converges in probability. Specifically, considering the uniform probability assumption, this implies that

$$\begin{aligned}\limsup_{N\to\infty} \frac{m_N}{N} &= \limsup_{N\to\infty} \frac{\#\{i \in [N] : I_n(i) \geq \epsilon\, d(X)\}}{N} \\ &= \limsup_{n\to\infty} \mathbb{P}(I_n \geq \epsilon\, d(X)) \\ &= \mathbb{P}(I_\infty \geq \epsilon\, d(X)) = d(X).\end{aligned}$$

It remains to prove that $\{\Phi_N\}$ is ϵ-REP. Let $S = \{i \in [N] : I_n(i) \geq \epsilon\, d(X)\}$ denote the selected rows to construct Φ_N and let $Z_1^N = H_N X_1^N$ be the full measurements. It is easy to check that $\Phi_N X_1^N = Z_S$. Also let $B_i = S \cap [i-1]$ denote all of the indices in S before i. We have

$$\begin{aligned}d(X_1^N|Z_S) &= d(Z_1^N|Z_S) = d(Z_{S^c}|Z_S) \\ &= \sum_{i\in S^c} d(Z_i|Z_{B_i}, Z_S) \\ &\leq \sum_{i\in S^c} d(Z_i|Z_1^{i-1}) \\ &= \sum_{i\in S^c} I_n(i) \leq N\epsilon\, d(X) = \epsilon\, d(X_1^N),\end{aligned}$$

which shows the ϵ-REP property for $\{\Phi_N\}$. ∎

Achievability proof for the discrete case: For the discrete case, the construction of the measurement family is very similar to the mixture case with the only difference that instead of using the erasure process corresponding to the RID, we use the discrete entropy function. More exactly, in the discrete case, assuming that $Z_1^N = H_N X_1^N$, we define the following process for $i \in [N]$, $I_n(i) = H(Z_i|Z_1^{i-1})$. In [15], using the conditional EPI result [16], the following was proved.

Lemma 1 ("Absorption phenomenon"). *$(I_n, \mathcal{F}_n, \mathbb{P})$ is a positive martingale converging to 0 almost surely.*

Similar to the mixture case, we again construct the family $\{\Phi_N\}$ by selecting those rows of the shuffled Hadamard matrix with I value greater than $\epsilon\, H(X_1)$.

Proof of Theorem 6: By a similar procedure, it is easy to show that $\{\Phi_N\}$ has zero measurement rate.

$$\begin{aligned}\limsup_{N\to\infty} \frac{m_N}{N} &= \limsup_{n\to\infty} \mathbb{P}(I_n \geq \epsilon H(X_1)) \\ &\leq \mathbb{P}(\limsup_{n\to\infty} I_n \geq \epsilon H(X_1)) \\ &= \mathbb{P}(I_\infty \geq \epsilon\, H(X_1)) = 0.\end{aligned}$$

Moreover, assuming that $S = \{i \in [N] : I_n(i) \geq \epsilon\, H(X_1)\}$ and $B_i = S \cap [i-1]$, we have

$$\begin{aligned}H(X_1^N|Z_S) &= H(Z_1^N|Z_S) = H(Z_{S^c}|Z_S) \\ &= \sum_{i\in S^c} H(Z_i|Z_{B_i}, Z_S) \\ &\leq \sum_{i\in S^c} H(Z_i|Z_1^{i-1}) \\ &= \sum_{i\in S^c} I_n(i) \leq N\,\epsilon\, H(X_1) = \epsilon\, H(X_1^N),\end{aligned}$$

which show the ϵ-REP property for $\{\Phi_N\}$. ∎

The last step is to prove Theorem 8, namely, to show that for a family of mixture distributions Π with strictly positive RID, there is a fixed measurement family $\{\Phi_N\}$ which is ϵ-REP for all of the distributions in Π with a measurement rate vector lying in the Rényi information region of of the family.

Proof of Theorem 8: The proof is simple considering the fact that the construction of the family $\{\Phi_N\}$ in the proof of Theorem 7 depends only on the erasure pattern. Also, the erasure pattern is independent of the shape of the distribution and only depends on its RID. Moreover, it can be shown that the erasure patterns for different value of δ are embedded in one another, namely, for $\delta > \delta'$, $I_n^\delta(i) \geq I_n^{\delta'}(i), i \in [N]$. Considering the method we use to construct the family $\{\Phi_N\}$, this implies that an ϵ-REP measurement family designed for a specific RID δ is ϵ-REP for any distribution with RID less than δ. Thus, if we design $\{\Phi_N\}$ for $\sup_{\pi\in\Pi} d(\pi)$, it will be ϵ-REP for any distribution in the family. ∎

Figure 1 shows the *absorption phenomenon* for a binary random variable with $\mathbb{P}(1) = p = 0.05$. Figure 2 shows the polarization of the RID for a random variable with RID 0.5.

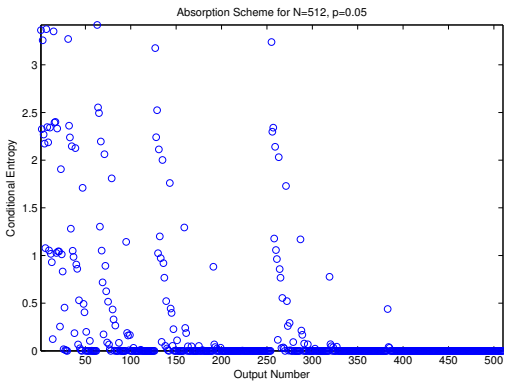

Fig. 1: Absorption pattern for $N = 512,\ p = 0.05$

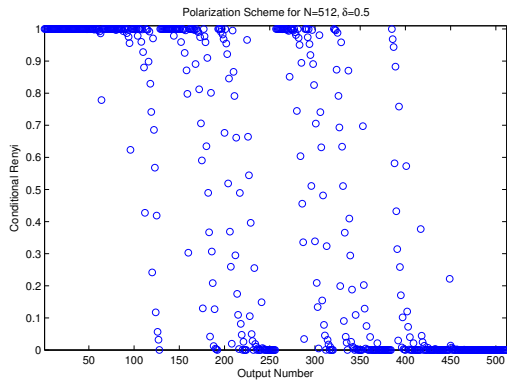

Fig. 2: Polarization of the RID for $N = 512, d(X) = 0.5$

D. Multi terminal A2A compression

In this section, we will give a brief overview of the techniques used to prove the achievability part. The proof of the converse part is given in Appendix B.

Acievability proof for the mixture case: The proof technique is very similar to the single terminal case. We will define the suitable erasure process and we will use it to construct the desired ϵ-REP measurement matrices for the multi terminal case. Let $\{(X_i, Y_i)\}_{i=1}^N$, $i \in [N]$, be a two-terminal memoryless source, where N is a power of two. Let $Z_1^N = H_N X_1^N$ and $W_1^N = H_N Y_1^N$. For $i \in [N]$, let us define $I_n(i) = d(Z_i|Z_1^{i-1})$ and $J_n(i) = d(W_i|W_1^{i-1}, Z_1^N)$. Using Theorem 4, we can show that I_n and J_n are erasure processes with initial values $d(X)$ and $d(Y|X)$ polarizing to $\{0, 1\}$.

The next step is to construct the two terminal measurement ensemble. Let $n \in \mathbb{N}$ and $N = 2^n$. We will construct Φ_N^x by selecting those rows of the Hadamard matrix, H_N, with $I_n(i) > \epsilon\, d(X)$. Similarly, Φ_N^y is constructed by selecting those rows of H_N with $J_n(i) > \epsilon\, d(Y|X)$. It remains to prove that the family $\{\Phi_N^x, \Phi_N^y\}$ labeled with N, a power of 2, and of dimension $m_N^x \times N$ and $m_N^y \times N$ is ϵ-REP with measurement rate $(d(X), d(Y|X))$. By this construction, we can achieve one of the corner points of the dominant face of the rate region. If we switch the role of X and Y we will get the other corner point $(d(X|Y), d(Y))$. One way to obtain any point on the dominant face is to use time sharing for the two family. However, it is also possible to use an explicit construction proposed in [22], which directly gives any point on the dominant face of the measurement rate region without any need to time sharing. We will just prove the achievability for the corner point $(d(X), d(Y|X))$.

Proof of Theorem 11: We first show that the family $\{\Phi_N^x, \Phi_N^y\}$ has measurement rate $(d(X), d(Y|X))$. Notice that the processes I_n^x, I_n^y converge almost surely thus, thay converge in probability. Specifically, considering the uniform probability assumption and using a similar technique as we used in the single terminal case, we get the following:

$$\begin{aligned}\limsup_{N\to\infty} \frac{m_N^x}{N} &= \limsup_{N\to\infty} \frac{\#\{i \in [N] : I_n^x(i) \geq \epsilon\, d(X)\}}{N} \\ &= \limsup_{n\to\infty} \mathbb{P}(I_n^x \geq \epsilon\, d(X)) \\ &= \mathbb{P}(I_\infty^x \geq \epsilon\, d(X)) = d(X).\end{aligned}$$

Similarly, we can show that $\limsup_{N\to\infty} \frac{m_N^y}{N} = d(Y|X)$.

It remains to prove that $\{\Phi_N^x, \Phi_N^y\}$ is ϵ-REP. Let $S_X = \{i \in [N] : I_n(i) \geq \epsilon\, d(X)\}$ and $S_Y = \{i \in [N] : J_n(i) \geq \epsilon\, d(Y|X)\}$ denote the selected rows to construct $\{\Phi_N^x, \Phi_N^y\}$ and let $Z_1^N = H_N X_1^N$ and $W_1^N = H_N Y_1^N$ be the full measurements for the x and the y terminal. Let $B_i^X = S_X^c \cap [1 : i-1]$ and $B_i^Y = S_Y^c \cap [1 : i-1]$ be the set of all indices in S_X^c and S_Y^c less than i. We have

$$\begin{aligned}d(X_1^N, Y_1^N|Z_{S_X}, W_{S_Y}) &= d(Z_1^N, W_1^N|Z_{S_X}, W_{S_Y}) \\ &\leq d(Z_1^N|Z_{S_X}) + d(W_1^N|Z_1^N, W_{S_Y}) \\ &\leq \sum_{i\in S_X^c} d(Z_i|Z_{B_i^X}, Z_{S_X}) \\ &+ \sum_{i\in S_Y^c} d(W_i|W_{B_i^Y}, W_{S_Y}, Z_1^N) \\ &\leq \sum_{i\in S_X^c} d(Z_i|Z_1^{i-1}) + \sum_{i\in S_Y^c} d(W_i|W_1^{i-1}, Z_1^N) \\ &\leq N\epsilon\, d(X) + N\epsilon\, d(Y|X) \\ &= \epsilon N d(X, Y) = \epsilon\, d(X_1^N, Y_1^N),\end{aligned}$$

which shows the ϵ-REP property for the two terminal measurement family $\{\Phi_N^x, \Phi_N^y\}$. ∎

Achievability proof for the discrete case: In the fully discrete case, the construction is very similar to the mixture case with the only difference that instead of using the RID, we will use the entropy. Similar to the single terminal case, we can prove the following.

Lemma 2. *$(I_n, \mathcal{F}_n)$ and $(J_n, \mathcal{F}_n)$ are positive martingale converging to 0 almost surely.*

We again construct the family $\{\Phi_N^x, \Phi_N^y\}$ by selecting those rows of H_N with $I_n > \epsilon H(X)$ and $J_n > \epsilon H(Y|X)$.

Proof of Theorem 10: Similar to the single terminal case, it is easy to show that $\{\Phi_N^x, \Phi_N^y\}$ has measurement rate $(0, 0)$.

It remains to prove that $\{\Phi_N^x, \Phi_N^y\}$ is ϵ-REP . Let $S_X = \{i \in [N] : I_n(i) \geq \epsilon\, H(X)\}$ and $S_Y = \{i \in [N] : J_n(i) \geq \epsilon\, H(Y|X)\}$ denote the selected rows to construct $\{\Phi_N^x, \Phi_N^y\}$ and let $Z_1^N = H_N X_1^N$ and $W_1^N = H_N Y_1^N$ be the full measurements for the X and the Y terminal. Let $B_i^X = S_X^c \cap [1 : i-1]$ and $B_i^Y = S_Y^c \cap [1 : i-1]$ be the set of all indices in S_X^c and S_Y^c less than i. We have the

following:

$$
\begin{aligned}
H(X_1^N, Y_1^N | Z_{S_X}, W_{S_Y}) &= H(Z_1^N, W_1^N | Z_{S_X}, W_{S_Y}) \\
&\le H(Z_1^N | Z_{S_X}) + H(W_1^N | Z_1^N, W_{S_Y}) \\
&\le \sum_{i \in S_X^c} H(Z_i | Z_{B_i^X}, Z_{S_X}) \\
&+ \sum_{i \in S_Y^c} H(W_i | W_{B_i^Y}, W_{S_Y}, Z_1^N) \\
&\le \sum_{i \in S_X^c} H(Z_i | Z_1^{i-1}) + \sum_{i \in S_Y^c} H(W_i | W_1^{i-1}, Z_1^N) \\
&\le N\epsilon\, H(X) + N\epsilon\, H(Y|X) \\
&= \epsilon N H(X,Y) = \epsilon\, H(X_1^N, Y_1^N),
\end{aligned}
$$

which shows the ϵ-REP property for the two terminal measurement family $\{\Phi_N^x, \Phi_N^y\}$. ∎

The last step is to prove Theorem 8, namely, to show for a family of mixture distributions Π, there is a fixed measurement family $\{\Phi_N^x, \Phi_N^y\}$, which is ϵ-REP for all of the distributions in Π with a measurement rate in the Rńyi information region of the family.

Proof of Theorem 12: The proof is simple considering the fact that the construction of the family $\{\Phi_N^x, \Phi_N^y\}$ in the proof of Theorem 11 depends only on the erasure pattern which is independent of the shape of the distribution and only depends on its RID. This implies that for any (ρ_x, ρ_y) in the Rényi information region of Π, the designed measurement family $\{\Phi_N^x, \Phi_N^y\}$ is ϵ-REP(Π). ∎

V. Numerical simulations

Up to now, we defined the notion of ϵ-REP for an ensemble of measurement matrices. This definition is what we call an "informational" characterization, in the sense that taking measurements by the ensemble potentially keeps more than $1-\epsilon$ ratio of the information of the source. Now, we can ask the natural question that weather this has some "operational" implication, in the sense that after having the linear measurements, is it possible to recover the source up to an acceptable distortion? In particular, is there a computationally feasible algorithm to do that?

To explain the operational view more, let us give an example from polar codes for binary source compression which has lots of similarities with what we have done. As shown in [20], for a binary memoryless source with $\mathbb{P}(0) = p$, for a large block length n, there is a matrix G_n, of dimension approximately equal to $nh_2(p) \times n$ such that the linear measurement of the source by this matrix over $\mathbb{F}_2$ faithfully captures all of the randomness of the source. This in its own only solves the encoding part of problem without directly addressing the decoding part, namely, it does not imply the existence of a decoder to recover the source from the measurements up to negligible distortion (error probability). Therefore, the operational picture is not complete yet. Fortunately, in the case of polar codes, the successive cancellation decoder (or other decoders proposed) fills up the gap and shows that the *informational* characterization implies the *operational* characterization.

For simulations, we use a unit variance sparse distribution $p_X(x) = (1-\delta)\delta_0(x) + \delta p_c(x)$, where $\delta_0(x)$ is the unit delta measure at point zero, p_c is the distribution of the continuous part and $\delta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ is the RID of the signal. We use the MSE (mean square error) as distortion measure. The simulations are done with the Hadamard matrix of order $N = 512$. To build the measurement matrix A, we select all of the rows of H_N with highest conditional RID, as stated in IV-C, until we get acceptable recovery distortion. Figure 3 shows the phase transition (PT) diagram for the ℓ_1-minimization algorithm. The simulations are done with 3 different distributions for p_c: Gaussian, Laplacian and Uniform. The acceptable recovery distortion is set to 0.01. The recovery is successful for the measurement rates above the plotted curves. The results show the insensitivity of the PT region to the distribution of the continuous components.

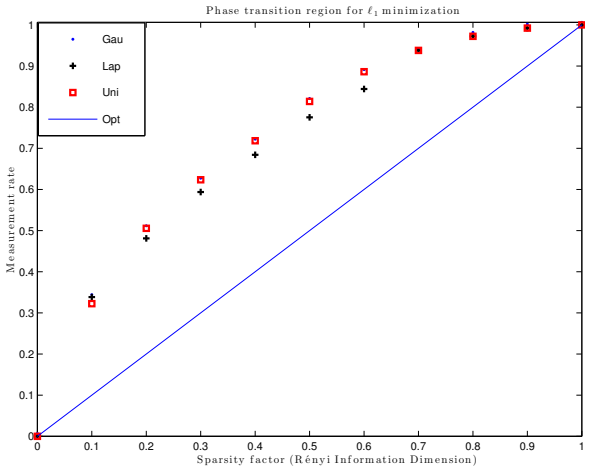

Fig. 3: PT diagram for ℓ_1-minimization

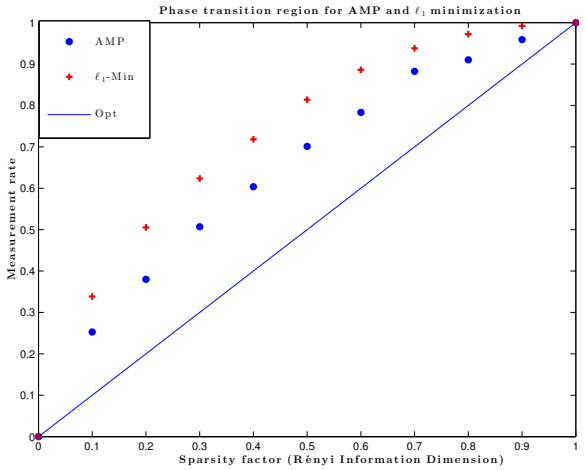

Fig. 4: PT diagram for AMP and ℓ_1-minimization

We also used the AMP algorithm to recover the signal, where for simplicity, we only did the simulations for the

Gaussian case for p_c. The AMP iteration is as follows:

$$z_t = y - A\hat{x}_t + \frac{1}{\gamma} z_{t-1} \langle \eta'_{t-1}(A^* z_{t-1} + \hat{x}_{t-1}) \rangle,$$
$$\hat{x}_{t+1} = \eta_t(A^* z_t + \hat{x}_t),$$

where $y = Ax$ is the linear measurements taken by A, γ is the measurement rate, $\langle a_1^n \rangle = \sum_{i=1}^n a_i/n$, $\eta_t(u) = (\eta_{t,1}(u_1), \ldots, \eta_{t,N}(u_N))$ and where $\eta_{t,i}(u_i) = \mathbb{E}\{X | u_i = X + \tau_t N\}$, with $N \sim \mathcal{N}(0,1)$ independent of the signal X and τ_t given by the state evolution equation for AMP, is the soft-thresholding function designed for the known distribution of X. For initialization, we use $\hat{x}_0 = 0$ and $z_0 = 0$. Figure 4 compares the PT diagram for AMP and ℓ_1-minimization. Although AMP, with the thresholding function η_t designed for the known distribution of the signal, performs better than ℓ_1-minimization, there is still a gap with the optimal line.

ACKNOWLEDGMENT

S. Haghighatshoar acknowledges Mr. Adel Javanmard for his helpful comments about the AMP algorithm. E. Abbe would like to thank Sergio Verdù for stimulating discussions on the Rényi information dimension.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] S. Kudekar and H.D. Pfister, "The effect of spatial coupling on compressive sensing," *In Proc. 48th Annual Allerton Conference,* 2010, pp. 347–353.

[4] F. Zhang and H.D. Pfister, "Verification decoding of high-rate LDPC codes with applications in compressed sensing," *IEEE Transactions on Inform. Theory*, vol. 58, pp. 5042–5058, Aug. 2012.

[5] A.G. Dimakis, R. Smarandache, P. Vontobel, "LDPC Codes for Compressed Sensing," *IEEE Transactions on Information Theory*, To appear.

[6] M. Akcakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling", *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.

[7] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proceedings of the 2008 IEEE International Symposium on Information Theory*, Toronto, Canada, Jul. 2008.

[8] S. Sarvotham, D. Baron, and R.G. Baraniuk, "Measurements and Bits: Compressed Sensing meets Information Theory," *Proceedings of the 44th Allerton Conference on Communication, Control, and Computing,* Monticello, IL, Sep. 2006.

[9] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007.

[10] W. Wang, M.J. Wainwright, K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Transactions on Information Theory*, Vol. 56, No. 6, pp. 2967–2979, Jun. 2010.

[11] D. Guo, D. Baron, and S. Shamai (Shitz), "A single-letter characterization of optimal noisy compressed sensing," *in Proceedings of the Forty-seventh Annual Allerton Conference on Communication,* Control, and Computing, Monticello, IL, Oct. 2009.

[12] Y. Wu and S. Verdú, "Rényi Information Dimension: Fundamental Limits of Almost Lossless Analog Compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3747, Aug. 2010.

[13] Y. Wu and S. Verdú, "Optimal Phase Transitions in Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6241-6263, Oct. 2012.

[14] D.L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," submitted to *IEEE Transactions Information Theory*, Dec. 2011.

[15] S. Haghighatshoar, E. Abbe, E. Telatar, "Adaptive sensing using deterministic partial Hadamard matrices," *IEEE International Symposium on Information Theory*, pp. 1842–1846, Jul. 2012.

[16] S. Haghighatshoar, E. Abbe, E. Telatar, "new entropy power inequality for integer-valued random variables," Jan. 2013, Available: http://arxiv.org/abs/1301.4185

[17] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical physicsbased reconstruction in compressed sensing," preprint, Nov. 2011. Available: http://arxiv.org/abs/1109.4424

[18] E. Abbe, "Universal source polarization and sparse recovery," IEEE Information Theory Workshop (ITW), Dublin, Aug. 2010.

[19] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions Inform. Theory*, vol. IT-55, pp. 3051–3073, Jul. 2009.

[20] E. Arıkan, "Source polarization," in *Proc. IEEE Int. Symp. Inform. Theory*, Austin, 2010.

[21] E. Arıkan and E. Telatar, "On the rate of channel polarization," *IEEE International Symposium on Information Theory*, pp. 1493–1495, Jul. 2009.

[22] E. Arıkan, "Polar coding for the Slepian-Wolf problem based on monotone chain rules," *IEEE Transactions Inform. Theory*, pp. 566 –570, Jul. 2012.

[23] D. Baron, M.F. Duarte, S. Sarvotham, M.B. Wakin, and R.G. Baraniuk, "An Information-Theoretic Approach to Distributed Compressed Sensing," *Proceedings of the 43rd Allerton Conference on Communication, Control, and Computing,* Monticello, IL, Sept. 2005.

[24] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources, *IEEE Transactions Inform. Theory*, vol. 19, pp. 471–480, Jul. 1973.

[25] E. J. Candès, T. Tao, "Decoding by linear programming", IEEE Transaction on Information Theory, vol.51, pp. 4203–4215, Dec. 2005.

[26] E. J. Candès, T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies", IEEE Transaction on Information Theory, vol. 52, pp. 5406–5425, Dec. 2006.

[27] A. Rényi,"On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, Mar. 1959.

APPENDIX A
PROOF OF THE CONVERSE PART FOR THE SINGLE TERMINAL

In this section, we will prove Theorem 5, which constitutes the converse part and puts a lower bound on the minimum number of linear measurements in order to to keep ϵ-REP property. We will prove the following lemmas which will be used repeatedly for other parts.

Lemma 3. *Assume that Φ is a full-rank matrix of dimension $m \times n$, for $m \leq n$, and $\det(\Phi\Phi^T) = 1$. Then, there exists $S \subset [n], |S| = m$ such that $|\det(\Phi_S)| \geq \frac{1}{\sqrt{\binom{n}{m}}} > 2^{-\frac{n}{2}}$.*

Proof: As $m \leq n$ from Cauchy-Binnet formula we have

$$1 = \det(\Phi\Phi^T) = \sum_{S \subset [n], |S|=m} \det(\Phi_S \Phi_S^T) = \sum_{S \subset [n], |S|=m} \det(\Phi_S)^2.$$

As all of the terms are positive, there must be a $S \subset [n]$ of size m such that $\det(\Phi_S)^2 \geq \frac{1}{\binom{n}{m}}$ which implies that $|\det(\Phi_S)| \geq \frac{1}{\sqrt{\binom{n}{m}}} > 2^{-\frac{n}{2}}$. ∎

Lemma 4. *Let X be a continuous random variable with finite differential entropy and let $D = [X]_q$. Suppose $\mathcal{O}$ is a random element for which the differential entropy and the RID of X given $\mathcal{O}$ are well-defined. Then, we have*

$$h(qX|D,\mathcal{O}) \leq 0,\ \lim_{q\to\infty} \frac{h(qX|D,\mathcal{O})}{\log_2(q)} = 0.$$

Proof: We have

$$h(qX|D,\mathcal{O}) = h(q(X-D)|D,\mathcal{O}) \leq h(q(X-D)).$$

We know that $0 \leq X - D \leq \frac{1}{q}$, which implies that $q(X-D)$ has a bounded support at most $[0,1]$. As the uniform distribution maximizes the differential entropy for a fixed support, we have $h(qX|D) \leq h(\mathcal{U}[0,1]) = 0$. We also have

$$\begin{aligned}\frac{h(qX|D,\mathcal{O})}{\log_2(q)} &= \frac{h(X|D,\mathcal{O}) + \log_2(q)}{\log_2(q)} \\ &= \frac{h(X|\mathcal{O}) - I(D;X|\mathcal{O}) + \log_2(q)}{\log_2(q)} \\ &= \frac{h(X|\mathcal{O}) - H(D|\mathcal{O}) + \log_2(q)}{\log_2(q)} \\ &= 1 - \frac{H(D|\mathcal{O})}{\log_2(q)} + \frac{h(X|\mathcal{O})}{\log_2(q)}.\end{aligned}$$

Given $\mathcal{O}$, X has a well-defined differential entropy, which implies that almost surely for all $\mathcal{O}$, X conditioned on $\mathcal{O}$ is a continuous random variable. Therefore,

$$\lim_{q\to\infty} \frac{H(D|\mathcal{O})}{\log_2(q)} = d(X|\mathcal{O}) = 1.$$

Taking the limit as q tends to the infinity we get the result. ■

Putting $\mathcal{O}$ equal to null in the Lemma 4, we get the following corollary.

Corollary 2. *Let X be a continuous random variable with a well-defined differential entropy and let $D = [X]_q$. Then $h(qX|D) \leq 0$ and $\lim_{q\to\infty} \frac{h(qX|D)}{\log_2(q)} = 0$.*

Lemma 5. *Let X_1^n be a sequence of i.i.d. continuous random variables and let $D_1^n = [X_1^n]_q$. Assume that Φ is a full-rank matrix of dimension $m \times n$ where $m \leq n$ and $\Phi\Phi^T = I_m$. Suppose $\mathcal{O}$ is a random element such that the differential entropy of ΦX_1^n given $\mathcal{O}$ is well-defined. Then $h(q\Phi X_1^n|D_1^n,\mathcal{O}) \doteq 0$.*

Proof: By Lemma 3, there is a $S \subset [n]$ of size m such that $|\det(\Phi_S)| \geq 2^{-\frac{n}{2}}$. Hence, we have

$$\begin{aligned}h(q\Phi X_1^n|D_1^n,\mathcal{O}) &= h(q\Phi_S X_S + \Phi_{S^c}X_{S^c}|D_1^n,\mathcal{O}) \\ &\geq h(q\Phi_S X_S + q\Phi_{S^c}X_{S^c}|D_1^n,\mathcal{O}) \\ &\geq h(q\Phi_S X_S + q\Phi_{S^c}X_{S^c}|X_{S^c},D_1^n,\mathcal{O}) \\ &= h(q\Phi_S X_S|X_{S^c},D_S,\mathcal{O}) \\ &= h(qX_S|D_S,X_{S^c},\mathcal{O}) + \log_2(|\det(\Phi_S)|) \\ &= \sum_{i\in S} h(qX_i|X_{B_i},X_{S^c},D_S,\mathcal{O}) - \frac{n}{2} \\ &= -\sum_{i\in S} |h(qX_i|X_{B_i},X_{S^c},D_S)| - \frac{n}{2} \\ &= -\sum_{i\in S} |h(qX_i|\mathcal{O}_i)| - \frac{n}{2} \doteq 0\end{aligned}$$

where $B_i = S \cap [i-1]$ and $\mathcal{O}_i = \{\mathcal{O}, X_{B_i}, X_{S^c}, D_{S\setminus\{i\}}\}$. The final result follows by applying Lemma 4. ■

Proof of Theorem 5: Without any loss of generality, we can assume that $\{\Phi_N\}$ is a full-rank family, otherwise, we can drop some of the rows of Φ_N and obtain an equivalent family with lower measurement rate. Also, we can assume that the rows of Φ_N are orthonormal. Otherwise, by Gram-Schmidt procedure, we can obtain an equivalent family with orthonormal rows. In other words, there is a lower triangular and invertible $m_N \times m_N$ matrix L_N such that $\tilde{\Phi}_N = L_N\Phi_N$ has orthonormal rows. As L_N is invertible, it results that

$$H(D_1^N|\tilde{\Phi}_N X_1^N) = H(D_1^N|L_N\Phi_N X_1^N) = H(D_1^N|\Phi_N X_1^N).$$

Thus the equivalent family $\{\tilde{\Phi}_N\}$ is also ϵ-REP and has orthonormal rows, namely, $\Phi_N\Phi_N^T = I_m$, where we again dropped the dependence of m on N. We also represent each $X_i, i \in [N]$ as $X_i = \Theta_i U_i + \bar{\Theta}_i V_i$.

From ϵ-REP assumption, for any $\eta > 0$ there is a $Q_1 \in \mathbb{N}$ such that for $q > Q_1$

$$\frac{I(D_1^N;\Phi_N X_1^N)}{N\log_2(q)} \geq \frac{H(D_1)(1-\epsilon-\eta)}{\log_2(q)}, \tag{10}$$

where $D_1^N = [X_1^N]_q$. As we are going to take the limit as q tends to infinity, we can drop the negligible terms. In other words, we have

$$\begin{aligned}\frac{I(D_1^N;\Phi_N X_1^N)}{N\log_2(q)} &\doteq \frac{I(D_1^N,\Theta_1^N;\Phi_N X_1^N)}{N\log_2(q)} \\ &\doteq \frac{I(D_1^N;\Phi_N X_1^N|\Theta_1^N)}{N\log_2(q)},\end{aligned} \tag{11}$$

where we used the fact that $I(\Theta_1^N;.) \leq NH(\Theta_1) \doteq 0$. For a specific realization θ_1^N, let $C_\theta = \{i \in [N] : \theta_i = 1\}$ and $\bar{C}_\theta = [N]\setminus C_\theta$ as introduced before. Then, we obtain

$$\begin{aligned}I(D_1^N;\Phi_N X_1^N|\theta_1^N) &= I(D_{C_\theta}, D_{\bar{C}_\theta};\Phi_{C_\theta}U_{C_\theta} + \Phi_{\bar{C}_\theta}V_{\bar{C}_\theta}) \\ &\doteq I(D_{C_\theta};\Phi_{C_\theta}U_{C_\theta} + \Phi_{\bar{C}_\theta}V_{\bar{C}_\theta}|D_{\bar{C}_\theta}) \\ &\doteq I(D_{C_\theta};\Phi_{C_\theta}U_{C_\theta} + \Phi_{\bar{C}_\theta}V_{\bar{C}_\theta}, V_{\bar{C}_\theta}|D_{\bar{C}_\theta}) \\ &\doteq I(D_{C_\theta};\Phi_{C_\theta}U_{C_\theta} + \Phi_{\bar{C}_\theta}V_{\bar{C}_\theta}|D_{\bar{C}_\theta}, V_{\bar{C}_\theta}) \\ &= I(D_{C_\theta};\Phi_{C_\theta}U_{C_\theta}),\end{aligned} \tag{12}$$

where $D_{C_\theta} = [U_{C_\theta}]_q$ and $D_{\bar{C}_\theta} = [V_{\bar{C}_\theta}]_q$ denote the component-wise quantization of U_{C_θ} and $V_{\bar{C}_\theta}$. We also used the fact that

$$H(D_{\bar{C}_\theta}) \le H(V_{\bar{C}_\theta}) \le NH(V_1) \doteq 0.$$

Let $D_u = [U_1]_q$ and $D_v = [V_1]_q$. We consider two cases: First, if $|C_\theta| \le m$, using (12), we have

$$I(D_{C_\theta}; \Phi_{C_\theta} U_{C_\theta}) \le mH(D_u) \triangleq B_1(\theta_1^N). \tag{13}$$

Second, if $|C_\theta| > m$, generally, Φ_{C_θ} is neither full-rank nor orthonormal. However, we can drop the redundant rows and by using the Gram-Scmidt procedure, we can create an equivalent orthonormal matrix $\tilde{\Phi}$ of dimension $m' \times |C_\theta|$ with $m' \le m < |C_\theta|$. Therefore, for this case we obtain

$$\begin{aligned} I(D_{C_\theta}; \Phi_{C_\theta} U_{C_\theta}) &= I(D_{C_\theta}; \tilde{\Phi} U_{C_\theta}) \\ &= h(\tilde{\Phi} U_{C_\theta}) - h(\tilde{\Phi} U_{C_\theta} | D_{C_\theta}) \\ &= h(\tilde{\Phi} U_{C_\theta}) - h(q\tilde{\Phi} U_{C_\theta} | D_{C_\theta}) + m' \log_2(q) \\ &\preceq \frac{m'}{2} \log_2(2\pi e \sigma_u^2) + m' \log_2(q) \\ &\doteq m' \log_2(q) \triangleq B_2(\theta_1^N), \end{aligned} \tag{14}$$

where σ_u^2 is the variance of U_1. We also used Lemma 5, $h(q\tilde{\Phi} U_{C_\theta} | D_{C_\theta}) \doteq 0$, and the fact that the Gaussian distribution maximizes the differential entropy for a given covariance matrix. Combining (13) and (14), and $m' \le m$ we obtain

$$I(D_1^N; \Phi_N X_1^N | \theta_1^N) \preceq m \max\{\log_2(q), H(D_u)\},$$

which implies that

$$I(D_1^N; \Phi_N X_1^N | \Theta_1^N) \preceq m \max\{\log_2(q), H(D_u)\}. \tag{15}$$

Moreover, from (10), (11) and (15), we get

$$\frac{m}{N} \max\{1, \frac{H(D_u)}{\log_2(q)}\} \preceq \frac{H(D_1)(1-\epsilon-\eta)}{\log_2(q)}.$$

Taking the limit as q tends to infinity, we obtain

$$\frac{m}{N} \ge \delta(1-\epsilon-\eta),$$

which implies that

$$\limsup_{N\to\infty} \frac{m}{N} \ge \delta(1-\epsilon-\eta).$$

As $\eta > 0$ is arbitrary, we get the result. ∎

APPENDIX B
PROOF OF THE CONVERSE PART FOR THE MULTI TERMINAL

This section is devoted to the proof of Theorem 9. This theorem puts constraints on the number of linear measurements we should take from different terminals in order to keep ϵ-REP property.

Proof of Theorem 9: From ϵ-REP property, we have

$$\begin{aligned} I([X_1^N]_q, [Y_1^N]_q; &\Phi_N^x X_1^N, \Phi_N^y Y_1^N) \\ &\ge (1-\epsilon) H([X_1^N]_q, [Y_1^N]_q). \end{aligned} \tag{16}$$

Similar to the Γ notation that we used in (8) for the representation for X_1^N and Y_1^N, we have

$$\begin{aligned} I([X_1^N]_q, [Y_1^N]_q; &\Phi_N^x X_1^N, \Phi_N^y Y_1^N) \\ &\doteq I([X_1^N]_q, [Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N, \Gamma_1^N) \\ &\doteq I([X_1^N]_q, [Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N | \Gamma_1^N) \end{aligned}$$

As Γ_1^N takes finitely many values, we can obtain the result for a specific realization γ_1^N and then take expectation over all possible realizations. For a specif realization γ_1^N, if some of the components of $\Phi_N^x X_1^N$ and $\Phi_N^y Y_1^N$ are discrete or they are linearly dependent we can drop them. With some abuse of notation, let $(\Phi_N^x X_1^N, \Phi_N^y Y_1^N)$ denote the remaining components which have will have dimension $r_N^x \times N$ and $r_N^y \times N$, where $r_N^x \le m_N^x$ and $r_N^y \le m_N^y$ depend on the specific realization γ_1^N.

$$\begin{aligned} &I_\gamma([X_1^N]_q, [Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N) \\ &= h_\gamma(\Phi_N^x X_1^N, \Phi_N^y Y_1^N) \qquad (17) \\ &- h_\gamma(\Phi_N^x X_1^N, \Phi_N^y Y_1^N | [X_1^N]_q, [Y_1^N]_q) \\ &\preceq -h_\gamma(\Phi_N^x X_1^N, \Phi_N^y Y_1^N | [X_1^N]_q, [Y_1^N]_q) \\ &\doteq -h_\gamma(q\Phi_N^x X_1^N, q\Phi_N^y Y_1^N | [X_1^N]_q, [Y_1^N]_q) \qquad (18) \\ &+ (r_N^x + r_N^y) \log_2(q) \\ &\preceq (m_N^x + m_N^y) \log_2(q), \end{aligned}$$

where in (17), we used the fact that $h_\gamma(\Phi_N^x X_1^N, \Phi_N^y Y_1^N)$ is upper bounded by the differential entropy of a Gaussian random vector with appropriate covariance matrix which vanishes in the limit as q tends to infinity. Also, in (18), we used Lemma 5 to show that $h_\gamma(q\Phi_N^x X_1^N, q\Phi_N^y Y_1^N | [X_1^N]_q, [Y_1^N]_q) \doteq 0$. Therefore, taking the expectation over Γ_1^N we obtain that

$$I([X_1^N]_q, [Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N) \preceq (m_N^x + m_N^y) \log_2(q).$$

We also have $H([X_1^N]_q, [Y_1^N]_q) \doteq Nd(X,Y) \log_2(q)$. Therefore, using (16) and taking the limit as q tends to infinity, we obtain

$$\rho_x + \rho_y = \frac{m_N^x}{N} + \frac{m_N^y}{N} \ge (1-\epsilon) d(X,Y).$$

To prove the other two inequalities, notice that

$$\begin{aligned} &I_\gamma([X_1^N]_q, [Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N) \\ &\quad = I_\gamma([Y_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N) \\ &\quad + I_\gamma([X_1^N]_q; \Phi_N^x X_1^N, \Phi_N^y Y_1^N | [Y_1^N]_q) \\ &\quad \le H_\gamma([Y_1^N]_q) + I_\gamma([X_1^N]_q; \Phi_N^x X_1^N | [Y_1^N]_q) \\ &\quad + I_\gamma([X_1^N]_q; \Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q). \end{aligned} \tag{19}$$

For the last term, $I_\gamma([X_1^N]_q; \Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q)$, we can again assume that we have dropped all of discrete and linearly dependent terms from $\Phi_N^y Y_1^N$ so that it has a well-defined differential entropy. Thus, we obtain

$$\begin{aligned} &I_\gamma([X_1^N]_q; \Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q) \\ &\quad = I_\gamma([X_1^N]_q; q\Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q) \\ &\quad = h_\gamma(q\Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q) \qquad (20) \\ &\quad - h_\gamma(q\Phi_N^y Y_1^N | \Phi_N^x X_1^N, [Y_1^N]_q, [X_1^N]_q). \qquad (21) \end{aligned}$$

Notice that, for the first term in (20),

$$h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q) \le h_\gamma(q\Phi_N^y(Y_1^N-[Y_1^N]_q)).$$

It is easy to see that the random vector $(Y_1^N-[Y_1^N]) \in [0,\frac{1}{q}]^{r_N^y}$ has a bounded support independent of q thus $h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q)$ has an upper bound independent of q. Therefore,

$$h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q) \preceq 0.$$

Using a similar argument for (21), we have

$$h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q,[X_1^N]_q) \preceq 0.$$

Assume L is a lower triangular invertible matrix, obtained through the Gram-Scmidt procedure, such that $L\Phi_N^y$ is an orthonormal matrix. Then applying Lemma 5, we obtain that

$$h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q) \succeq -\log_2(|\det(L)|) \succeq 0,$$
$$h_\gamma(q\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q,[X_1^N]_q) \succeq -\log_2(|\det(L)|) \succeq 0.$$

This implies that $I_\gamma([X_1^N]_q;\Phi_N^y Y_1^N|\Phi_N^x X_1^N,[Y_1^N]_q) \doteq 0$. Thus, from (19), we obtain

$$\begin{aligned} &I_\gamma([X_1^N]_q,[Y_1^N]_q;\Phi_N^x X_1^N,\Phi_N^y Y_1^N) \\ &\quad\preceq H_\gamma([Y_1^N]_q) + I_\gamma([X_1^N]_q;\Phi_N^x X_1^N|[Y_1^N]_q). \end{aligned}$$

Again if $\Phi_N^x X_1^N$ has discrete components or if some of the components are linearly dependent or can be predicted from $[Y_1^N]_q$ we can drop them. With some abuse of notation let $\Phi_N^x X_1^N$ denote the resulting random vector of dimension $r_N^x \le m_N^x$. We have

$$\begin{aligned} &I_\gamma([X_1^N]_q;\Phi_N^x X_1^N|[Y_1^N]_q) \\ &\quad= h_\gamma(\Phi_N^x X_1^N|[Y_1^N]_q) - h_\gamma(\Phi_N^x X_1^N|[Y_1^N]_q,[X_1^N]_q) \\ &\quad\doteq -h_\gamma(\Phi_N^x X_1^N|[Y_1^N]_q,[X_1^N]_q) \\ &\quad\doteq -h_\gamma(q\Phi_N^x X_1^N|[Y_1^N]_q,[X_1^N]_q) + r_N^x\log_2(q) \\ &\quad\doteq r_N^x\log_2(q) \preceq m_N^x\log_2(q), \end{aligned} \tag{22}$$

where we used the fact that $h_\gamma(\Phi_N^x X_1^N|[Y_1^N]_q) \doteq 0$ and from Lemma 5, $h_\gamma(q\Phi_N^x X_1^N|[Y_1^N]_q,[X_1^N]_q) \doteq 0$. Therefore, taking the expectation over Γ_1^N and using (16) and (22), we obtain

$$\begin{aligned} m_N^x\log_2(q) &\succeq (1-\epsilon)H([X_1^N]_q,[Y_1^N]_q|\Gamma_1^N) - H([Y_1^N]_q|\Gamma_1^N), \\ &\doteq (1-\epsilon)H([X_1^N]_q,[Y_1^N]_q) - H([Y_1^N]_q), \end{aligned}$$

which implies that $\frac{m_N^x}{N} \ge d(X|Y) - \epsilon d(X,Y)$. Therefore, taking the limit as N tends to infinity, we get

$$\rho_x = \limsup_{N\to\infty}\frac{m_N^x}{N} \ge d(X|Y) - \epsilon d(X,Y).$$

The last inequality in the theorem follows by symmetry. ■